Beyond fat-trees without antennae, mirrors, and disco-balls

Simon Kassing, Asaf Valadarsky, Gal Shahaf,
Michael Schapira, Ankit Singla

ETH zürich

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

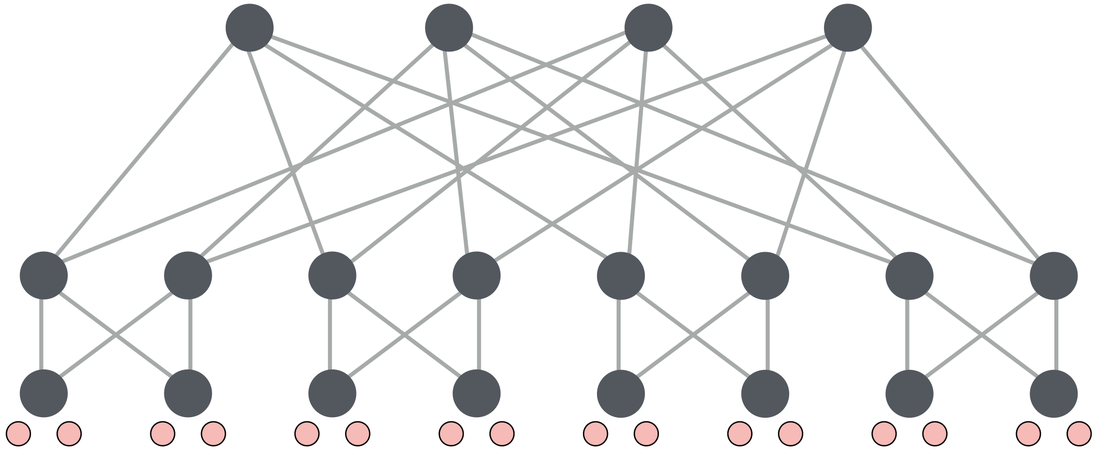
Skewed traffic within data centers

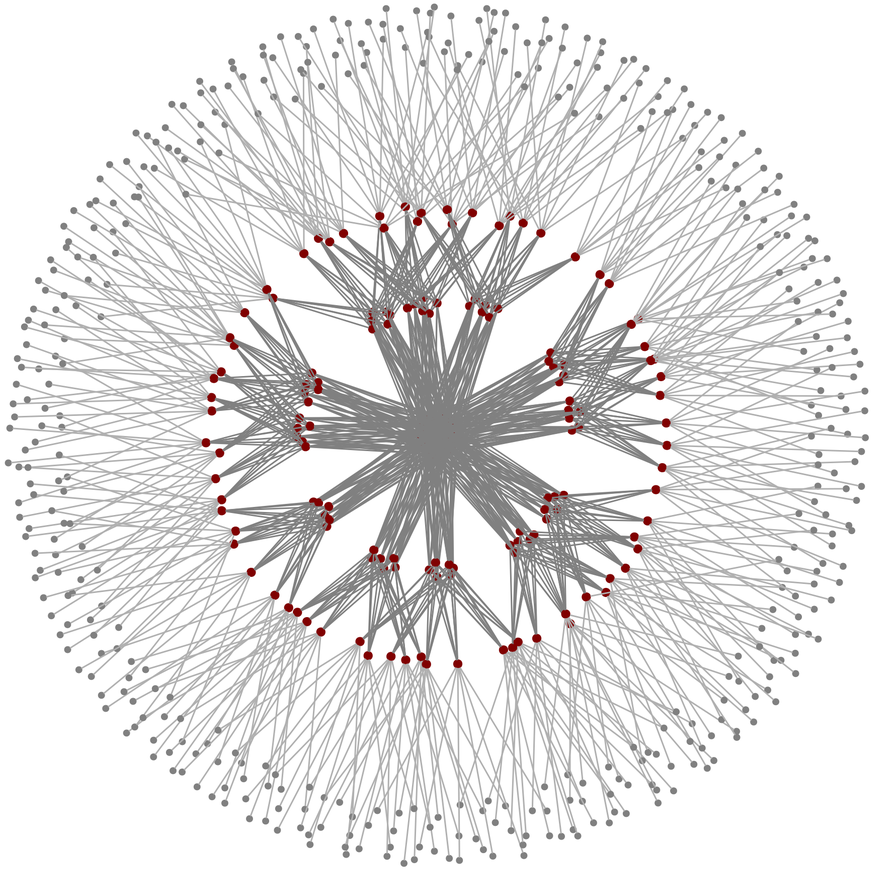
[Google]

Skewed traffic within data centers

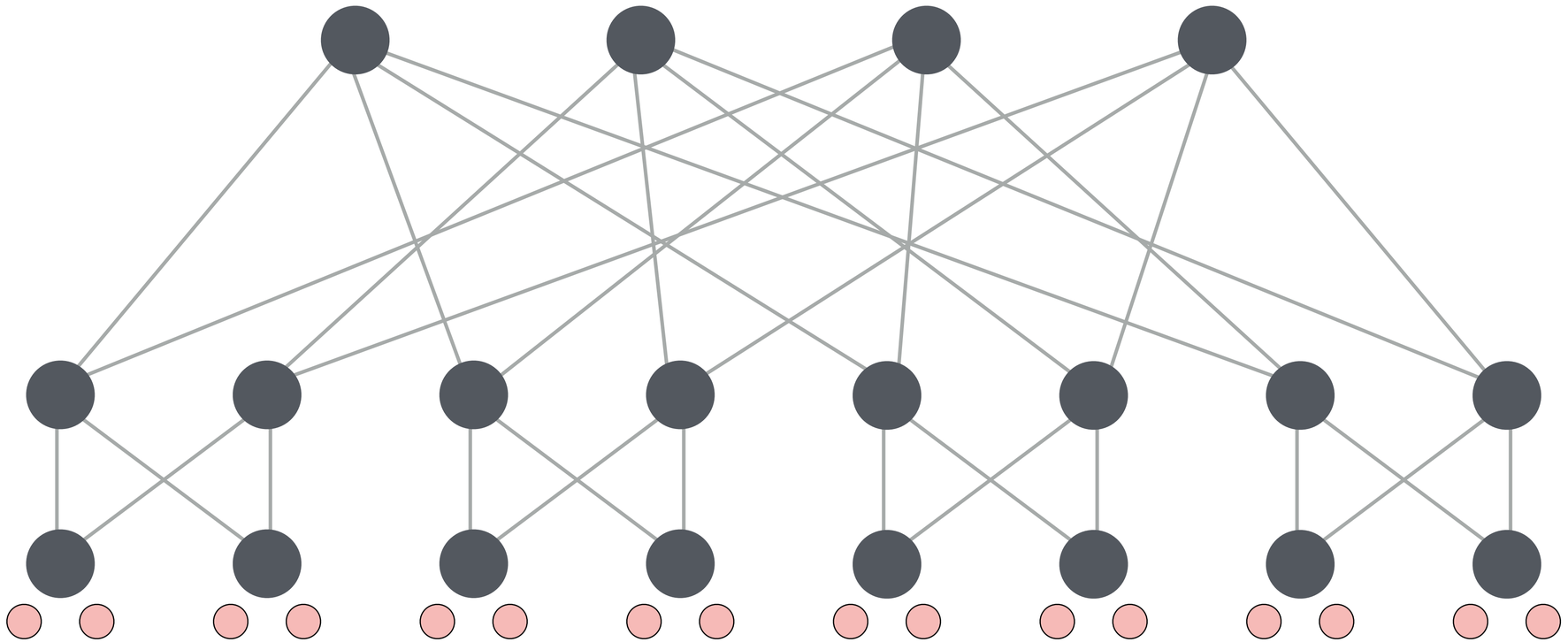
[Google]

All-to-all non-blocking connectivity is expensive

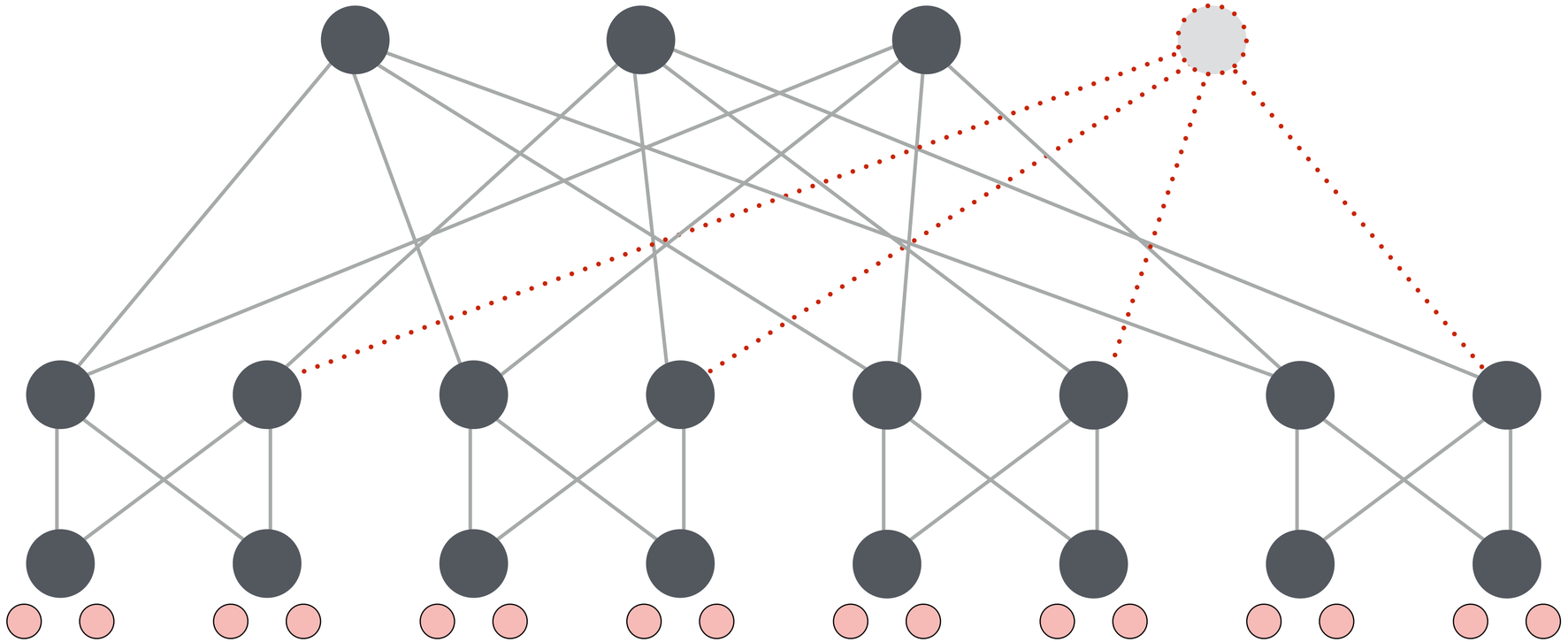




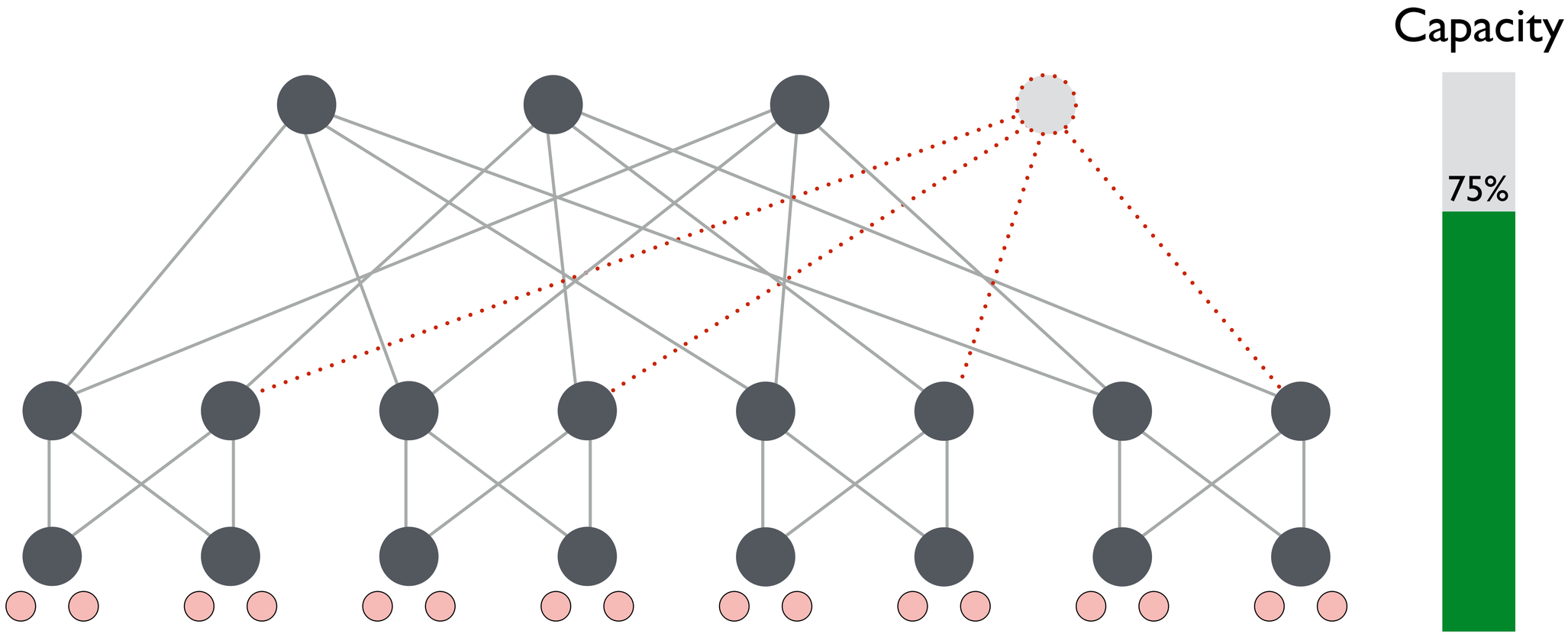
Oversubscribed fat-trees



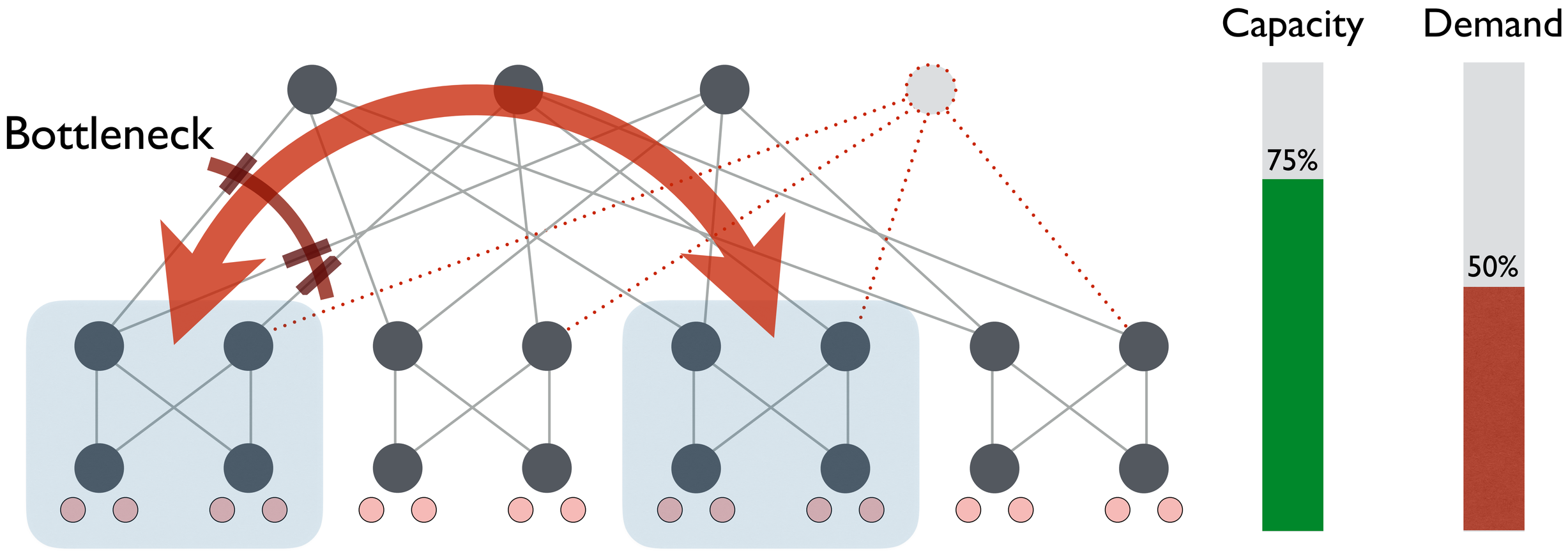
Oversubscribed fat-trees



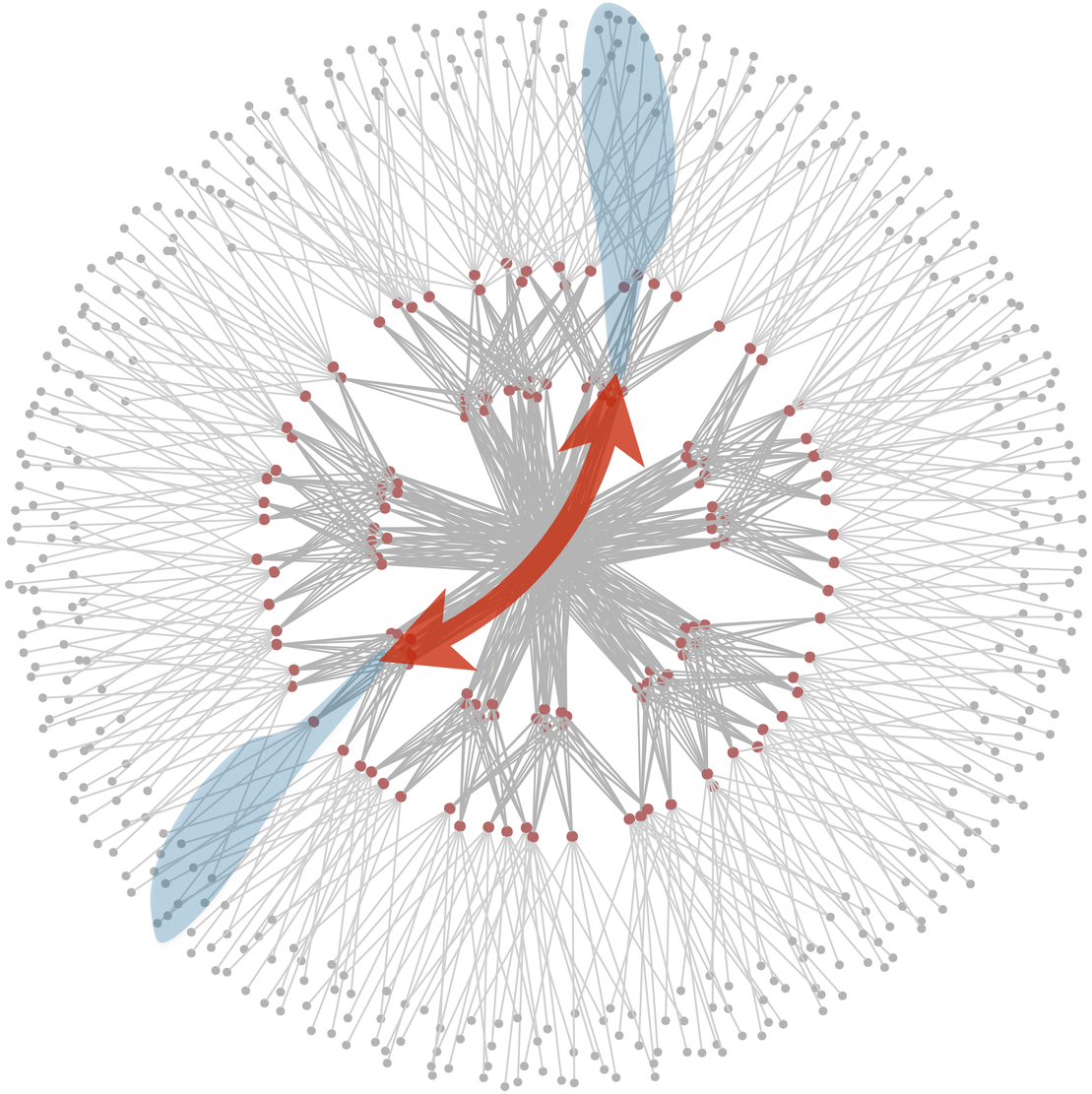
Oversubscribed fat-trees



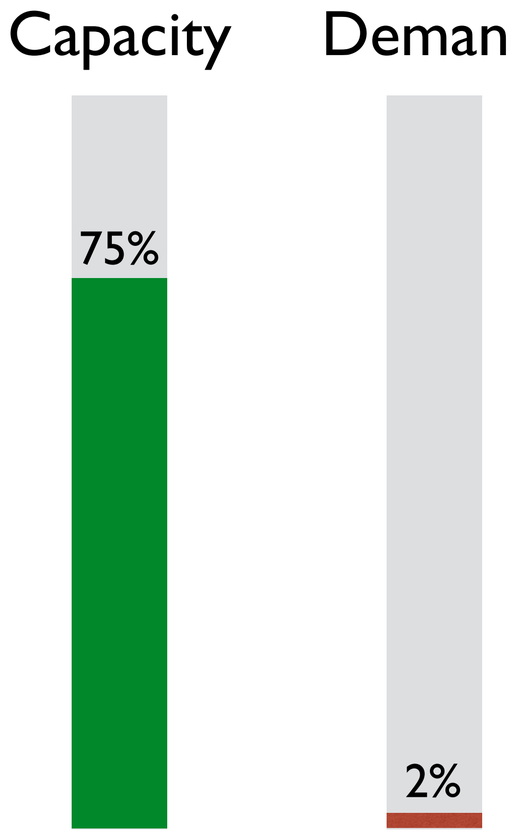
Oversubscribed fat-trees



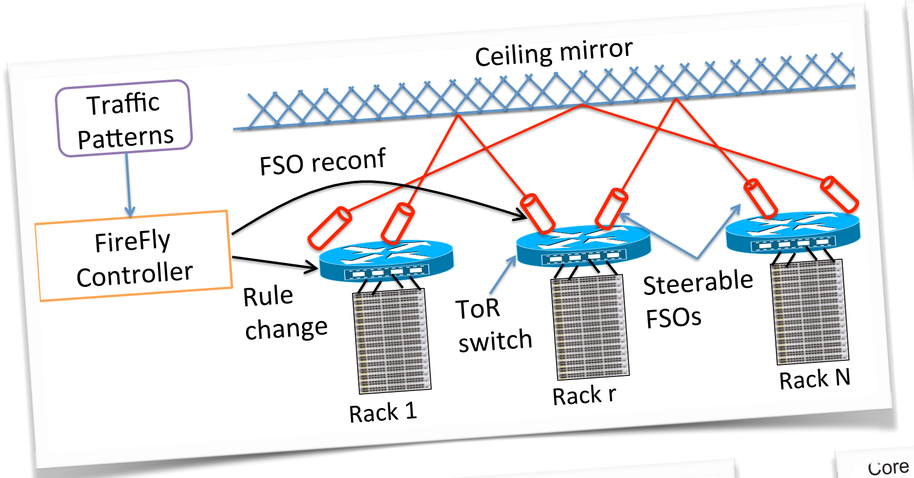
Oversubscribed fat-trees: A tragedy …

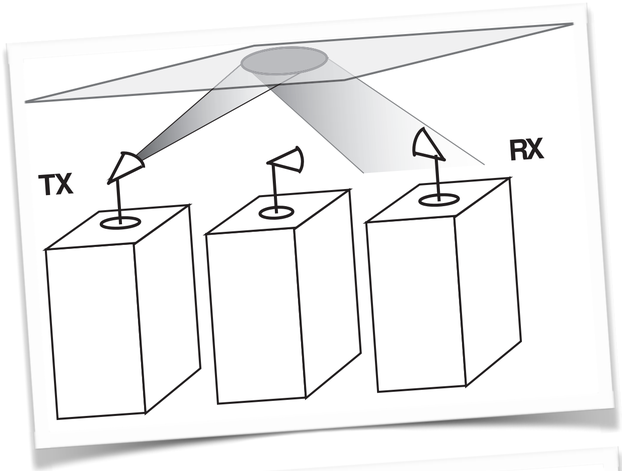


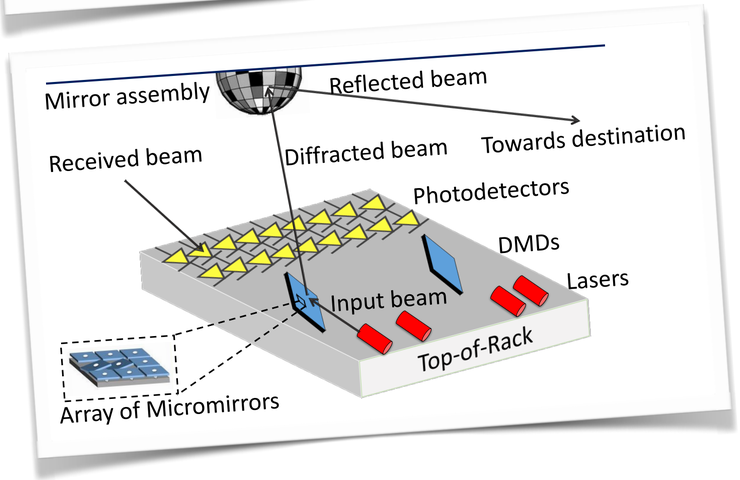
k = 96

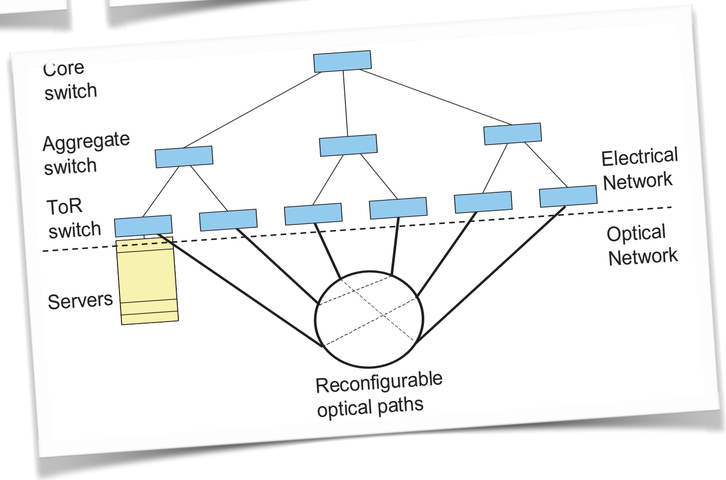


Dynamically set up network connections!









- OFC '09 Glick et al.
- SIGCOMM '10 Wang et al.
- SIGCOMM '10 Farrington et al.
- SIGCOMM '11 Halperin et al.
- NSDI '12 Chen et al.
- SIGCOMM '12 Zhou et al.
- SIGCOMM '13 Porter et al.
- SIGCOMM '14 Liu et al.
- SIGCOMM '14 Hamedazimi et al.
- SIGCOMM '16 Ghobadi et al.
- NSDI '17 Chen et al.

Set up network connections **on the fly!**

Advantage:
Gained the ability to move links around

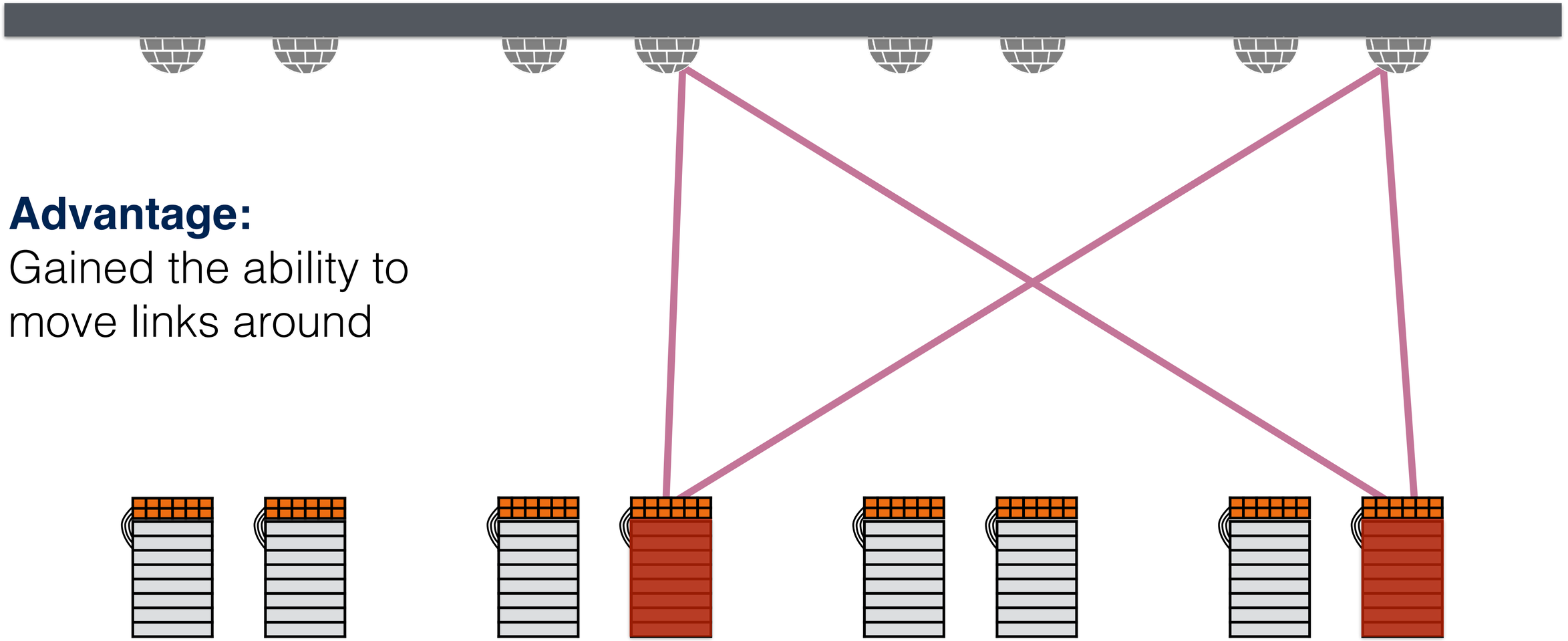
ProjecToR: Ghobadi, Monia, et al. "Projector: Agile reconfigurable data center interconnect." *SIGCOMM (*2016).

Set up network connections **on the fly!**

Advantage:
Gained the ability to move links around

ProjecToR: Ghobadi, Monia, et al. "Projector: Agile reconfigurable data center interconnect." *SIGCOMM (*2016).

Set up network connections **on the fly!**



Advantage:
Gained the ability to move links around

ProjecToR: Ghobadi, Monia, et al. "Projector: Agile reconfigurable data center interconnect." *SIGCOMM (*2016).

Engineering challenges facing dynamic topologies


Spatial planning and organisation?


Environmental factors?


Lack of operational experience?


Device packaging?


Monitoring and debugging?


Reliability and lifetime of devices?


Unknown unknowns?

Foundational questions

Foundational questions

1. Rigorous benchmarks?

   Fat-trees are the easiest baseline — ideally inflexible!

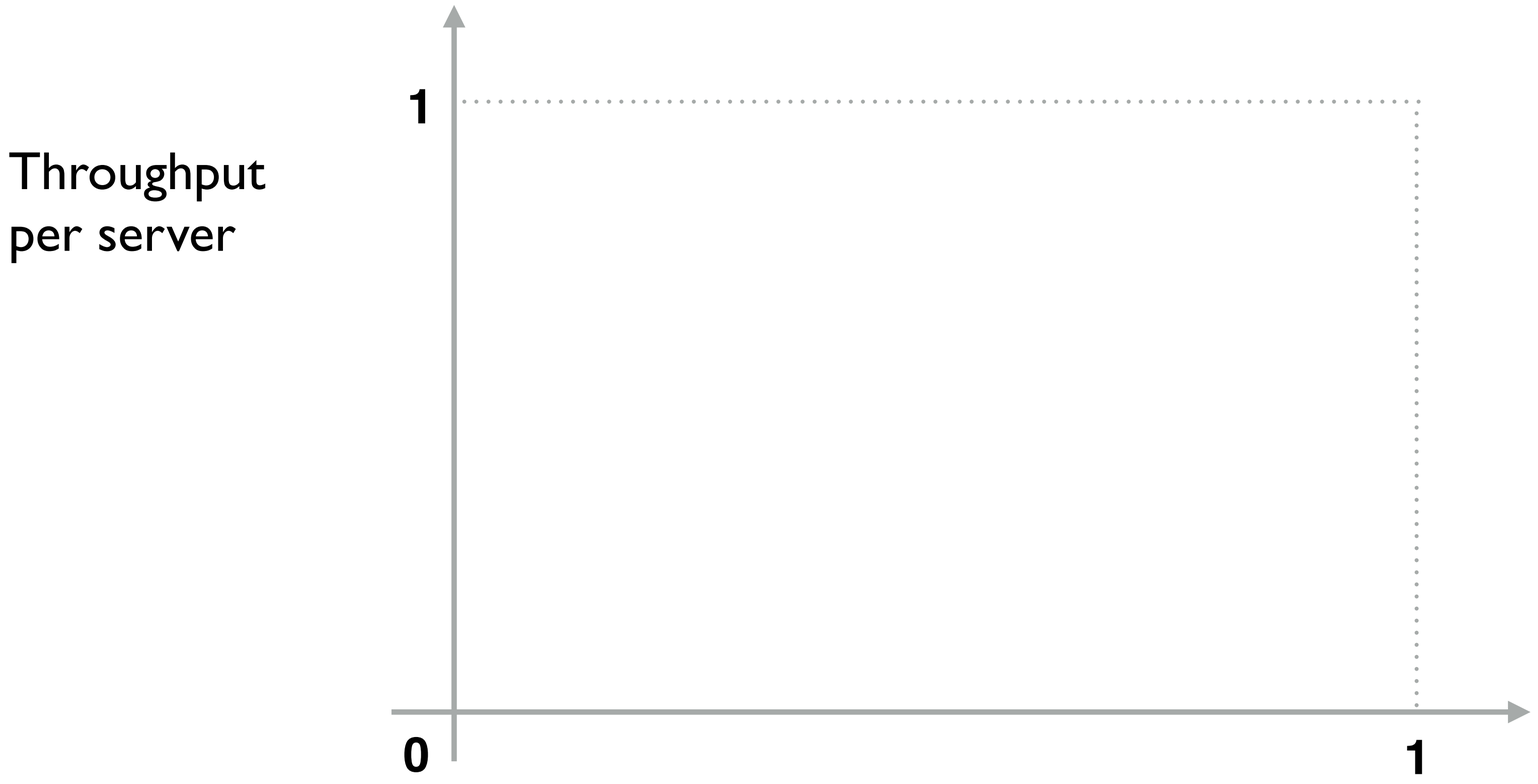
Foundational questions

1. Rigorous benchmarks?

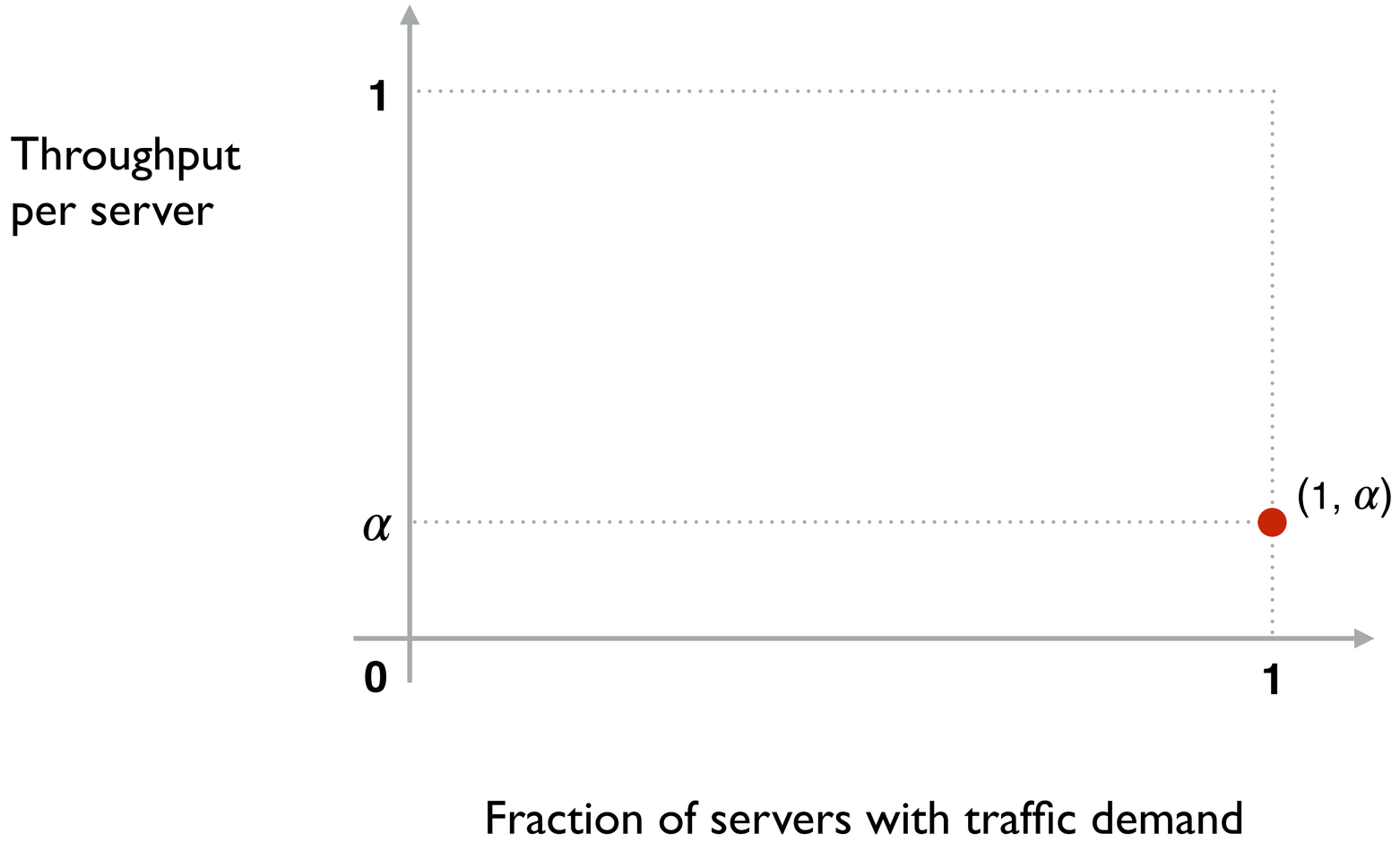
   Fat-trees are the easiest baseline — ideally inflexible!

2. What is the utility of dynamic links?

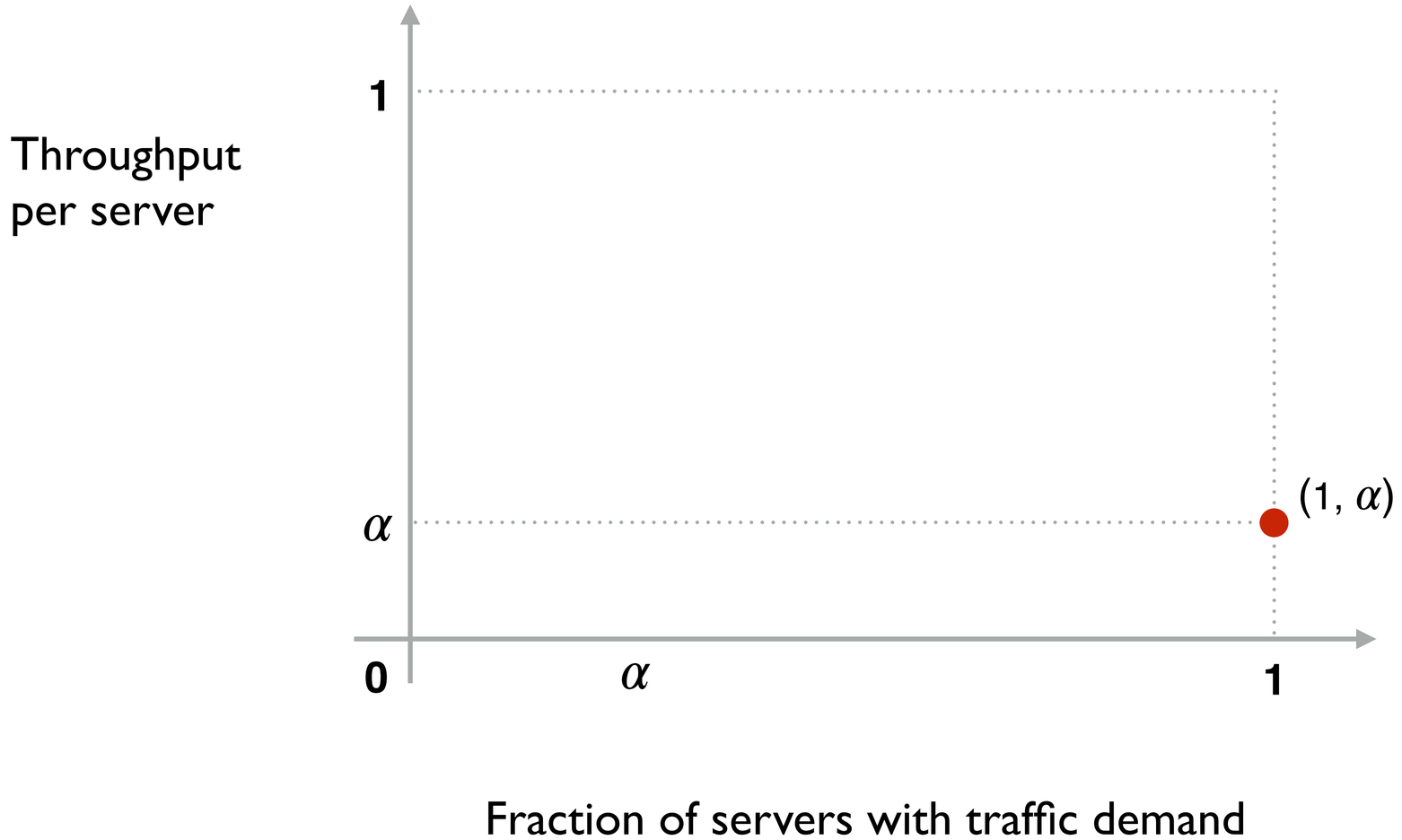
Ideally flexible network



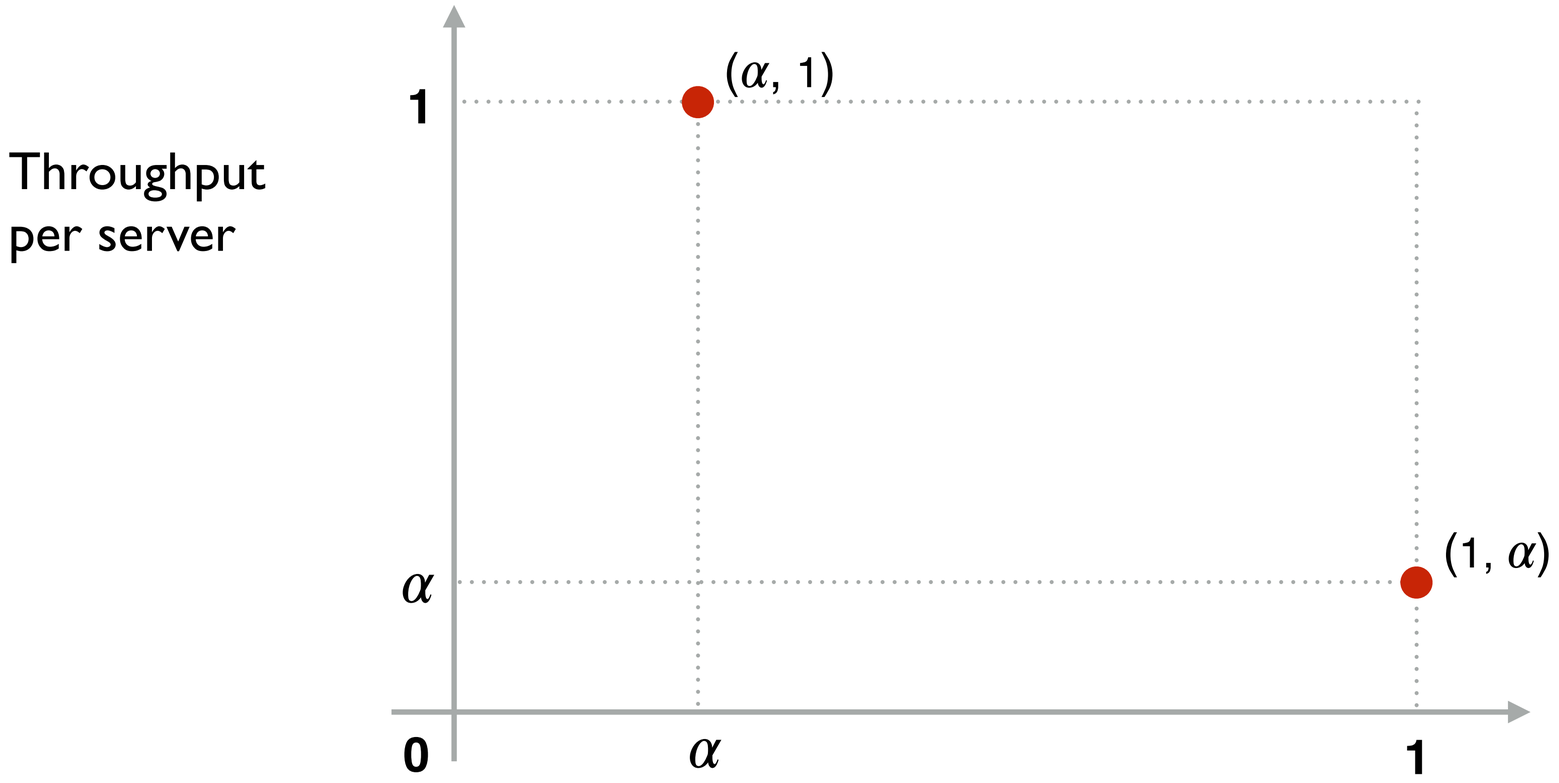
Ideally flexible network



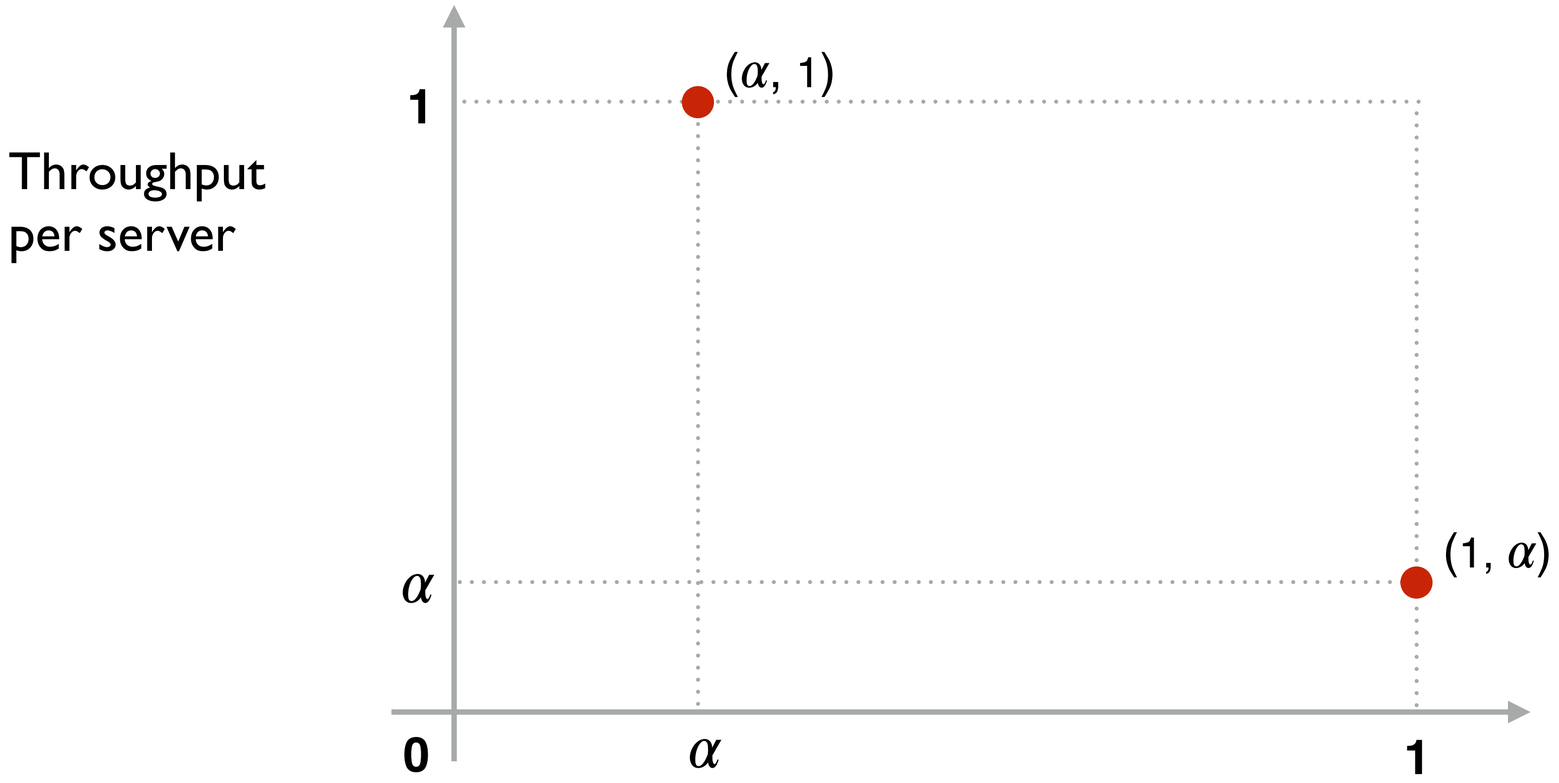
Ideally flexible network



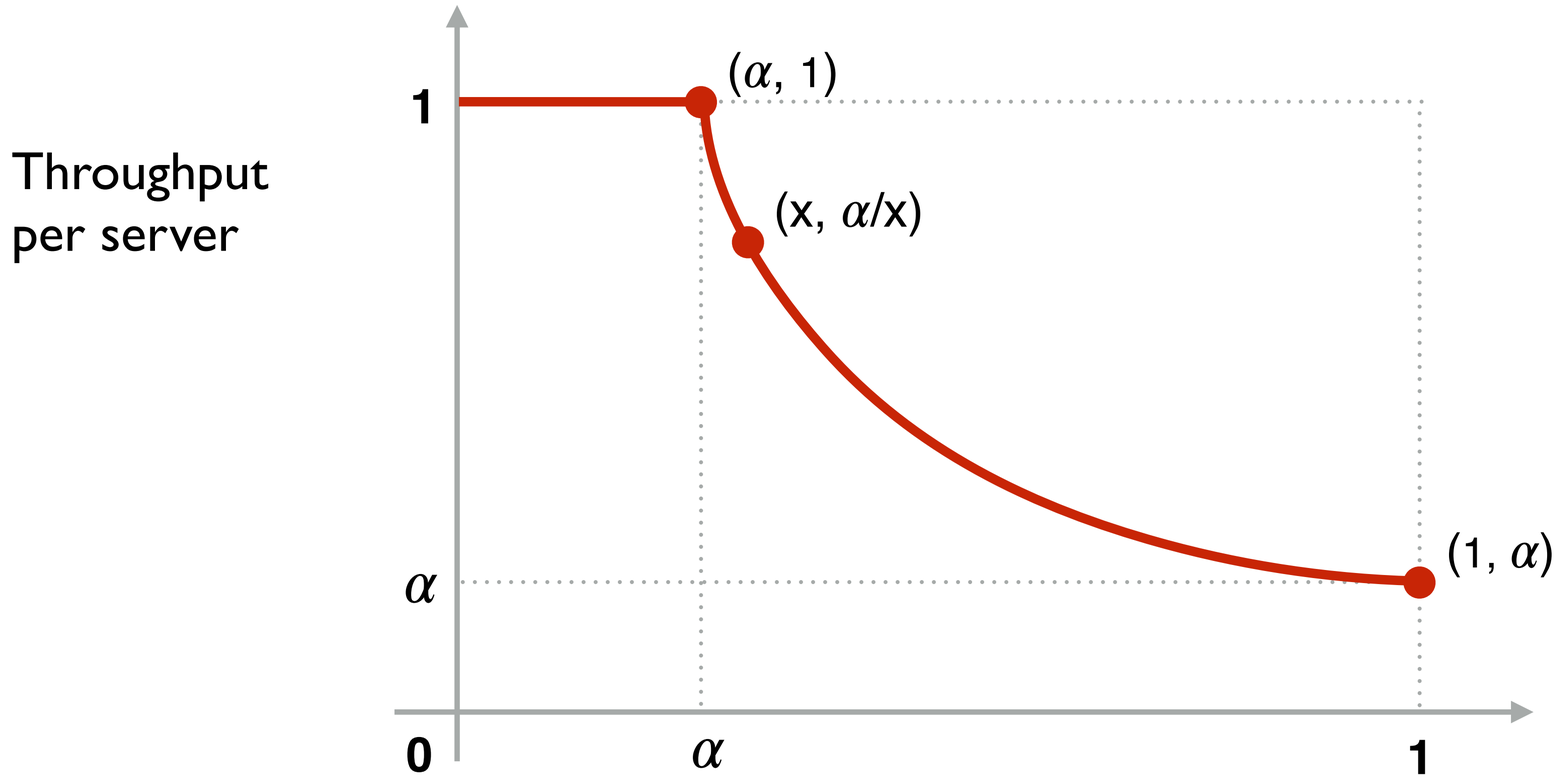
Ideally flexible network



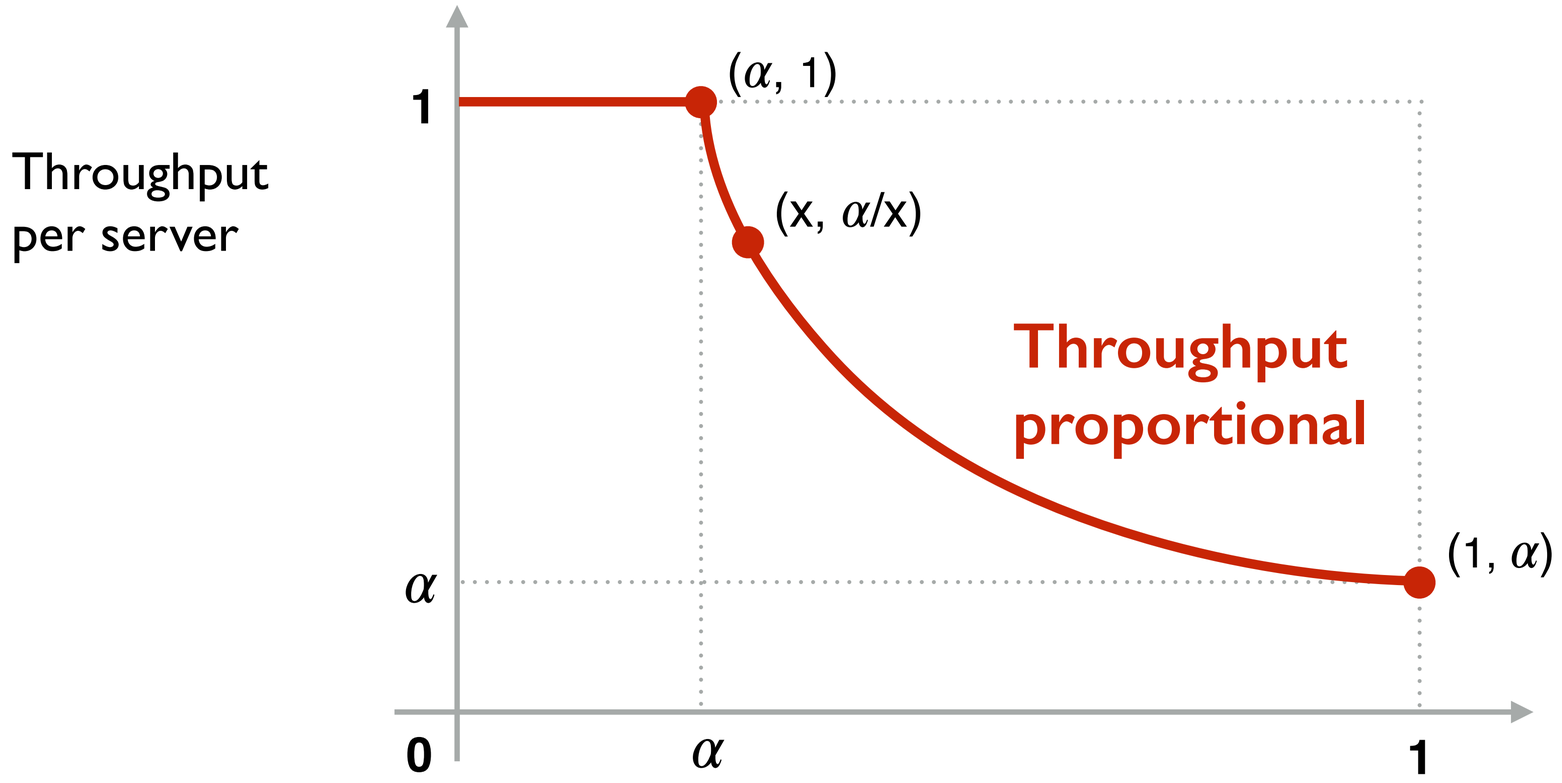
Ideally flexible network



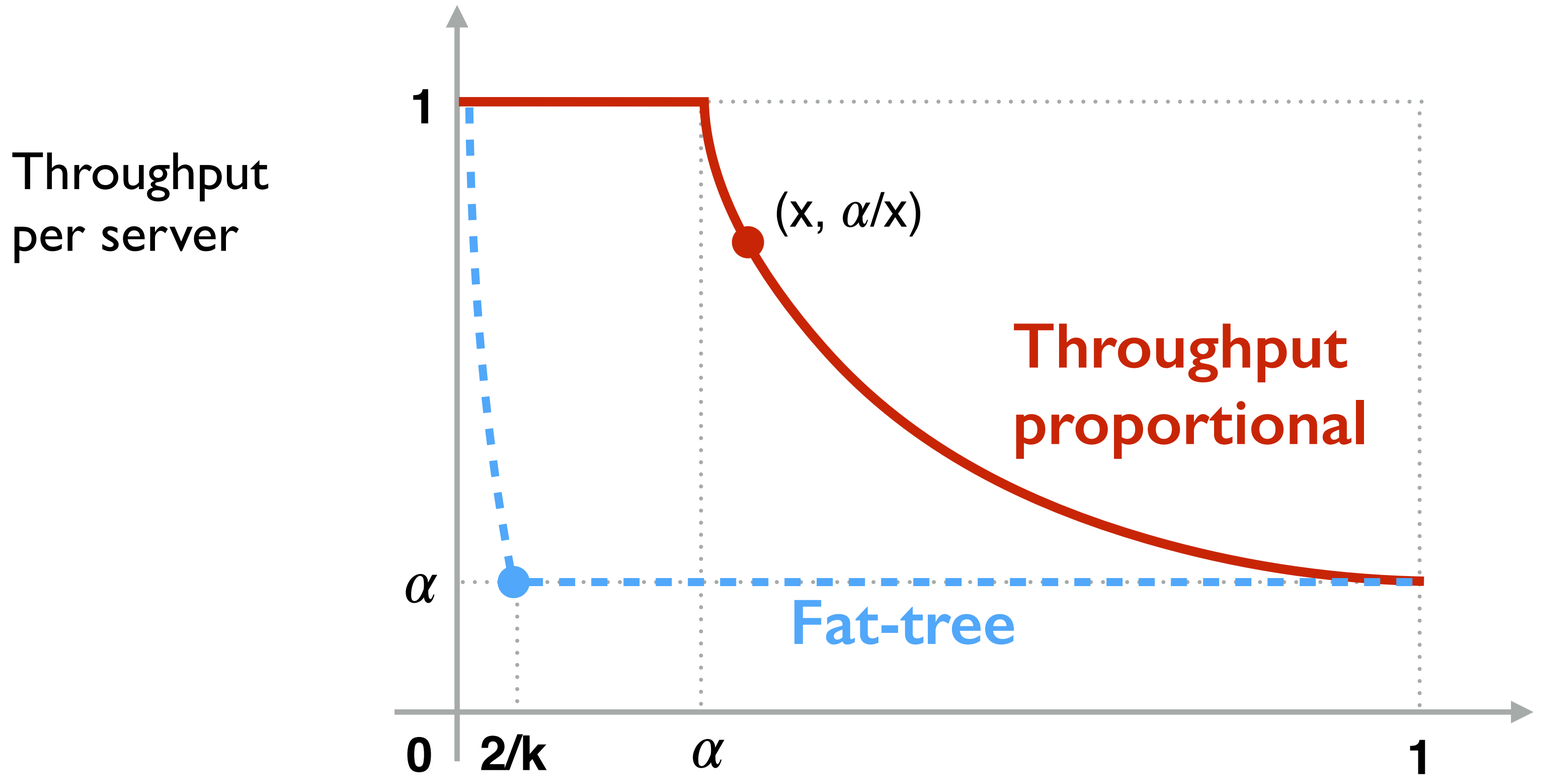
Ideally flexible network



Ideally flexible network



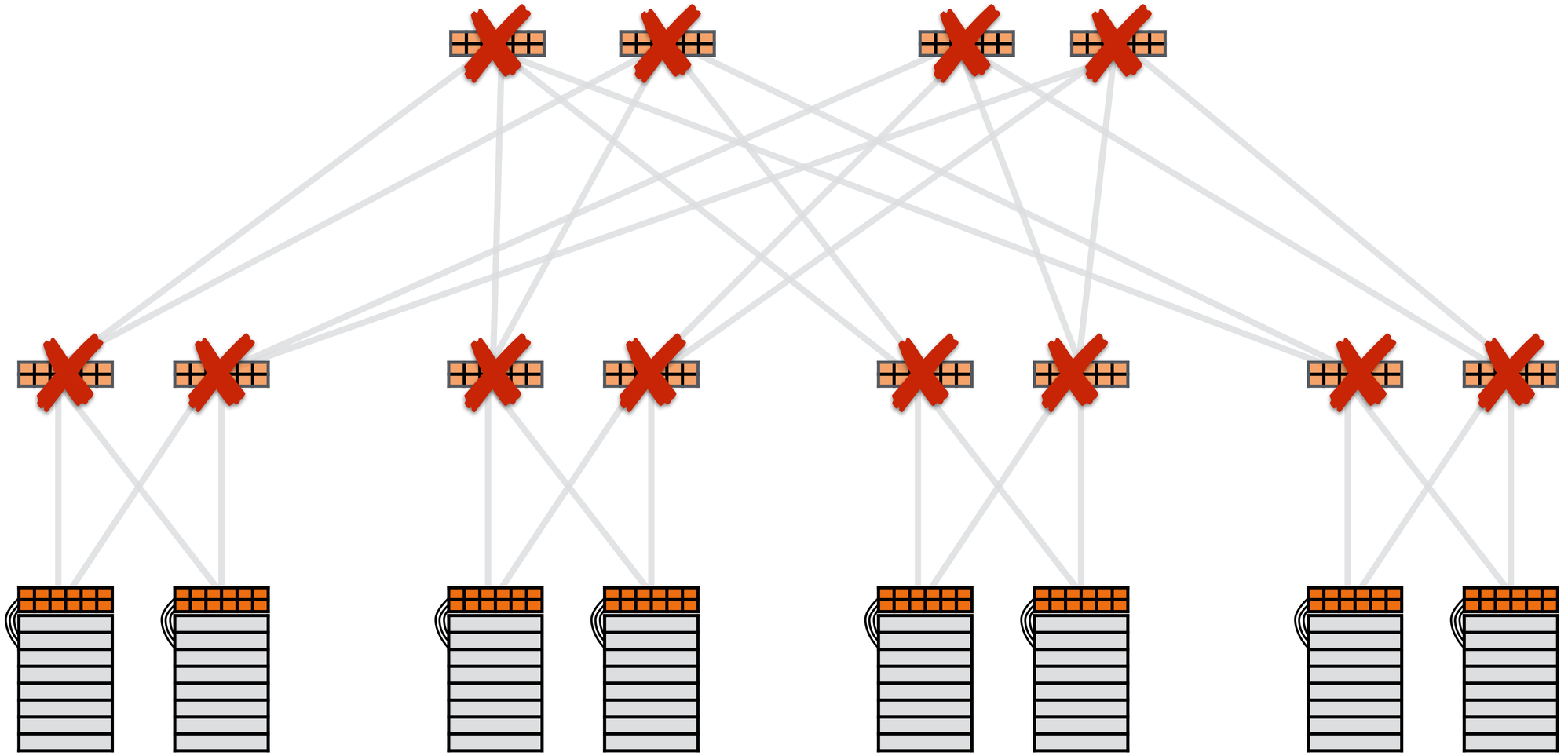
Fat-trees: ideally **in**flexible



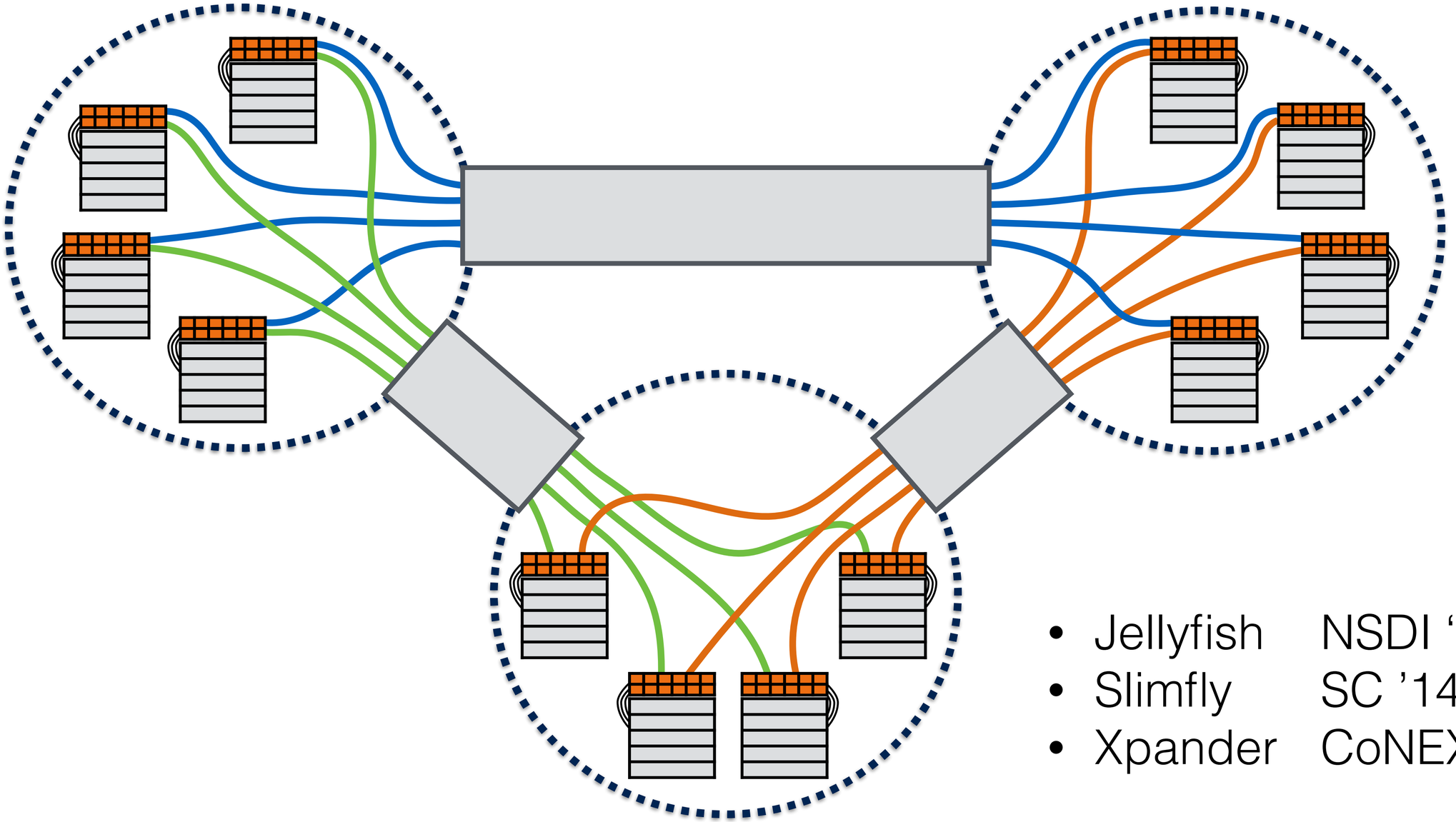
Near-optimal expander networks

Static but flexible

Instead of **rigid, layered connectivity**…

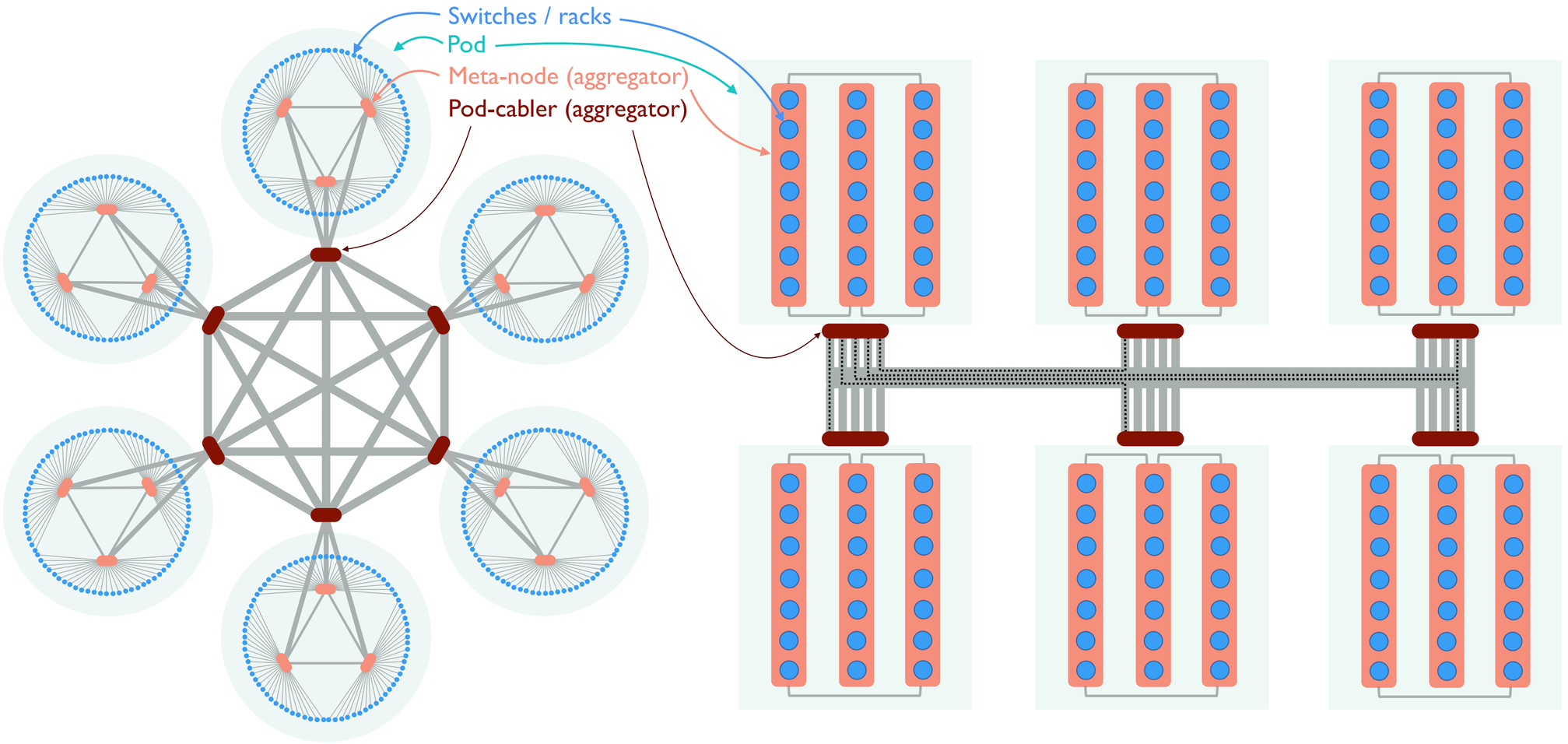


Expander-based data centers



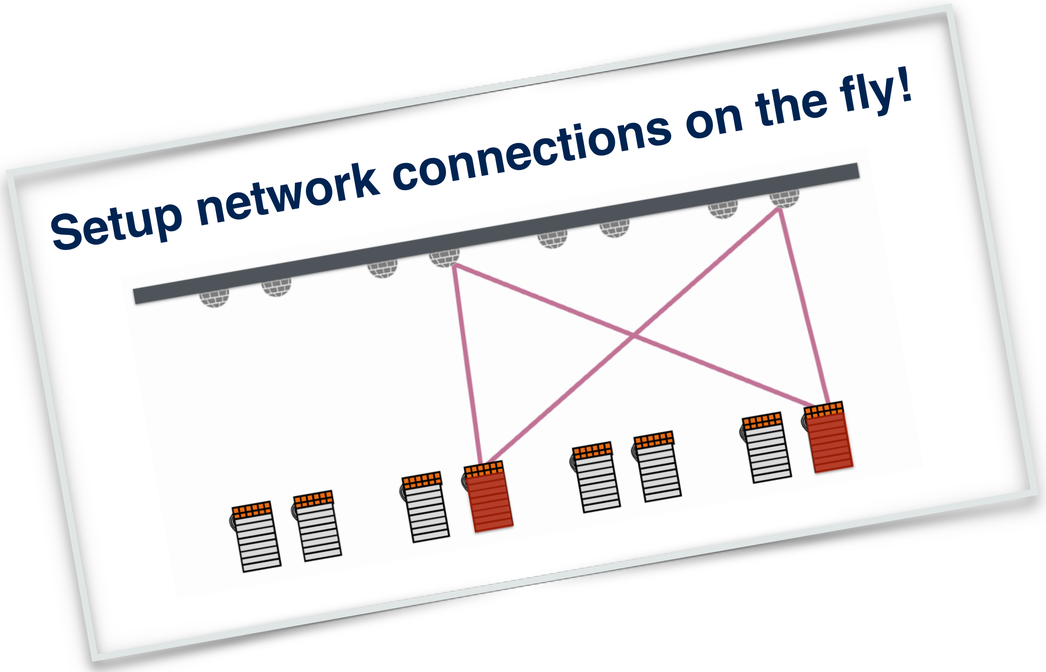
- Jellyfish NSDI ‘12
- Slimfly SC ’14
- Xpander CoNEXT ‘16

Xpander: deterministic wiring-friendly expander-based data center

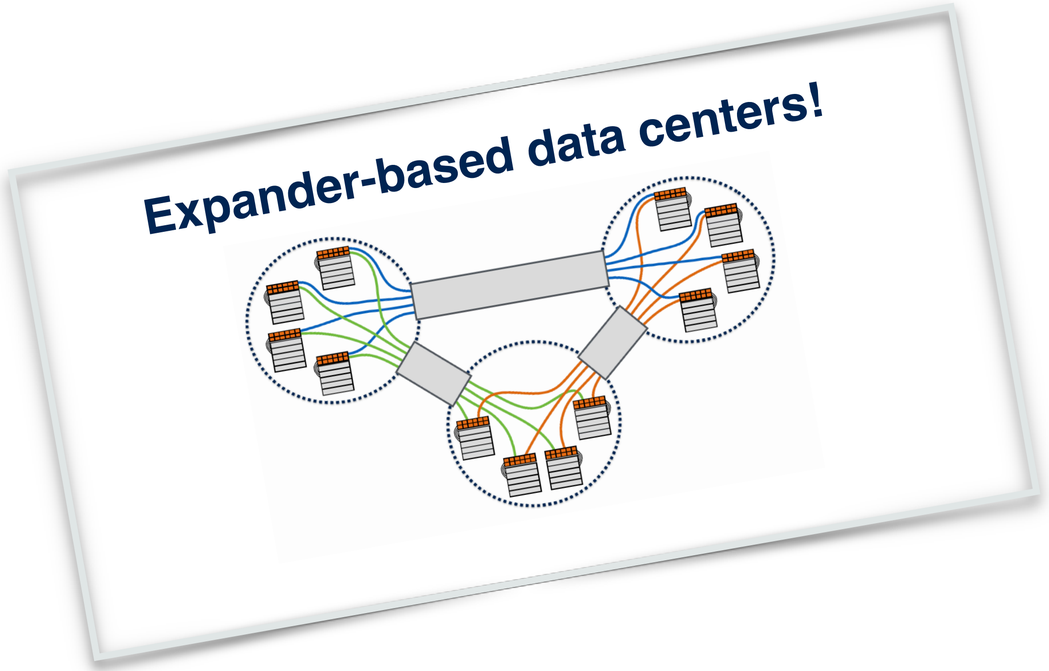


Valadarsky, Asaf, et al. "Xpander: Towards Optimal-Performance Datacenters." *CoNEXT*. 2016.

A fundamental question…

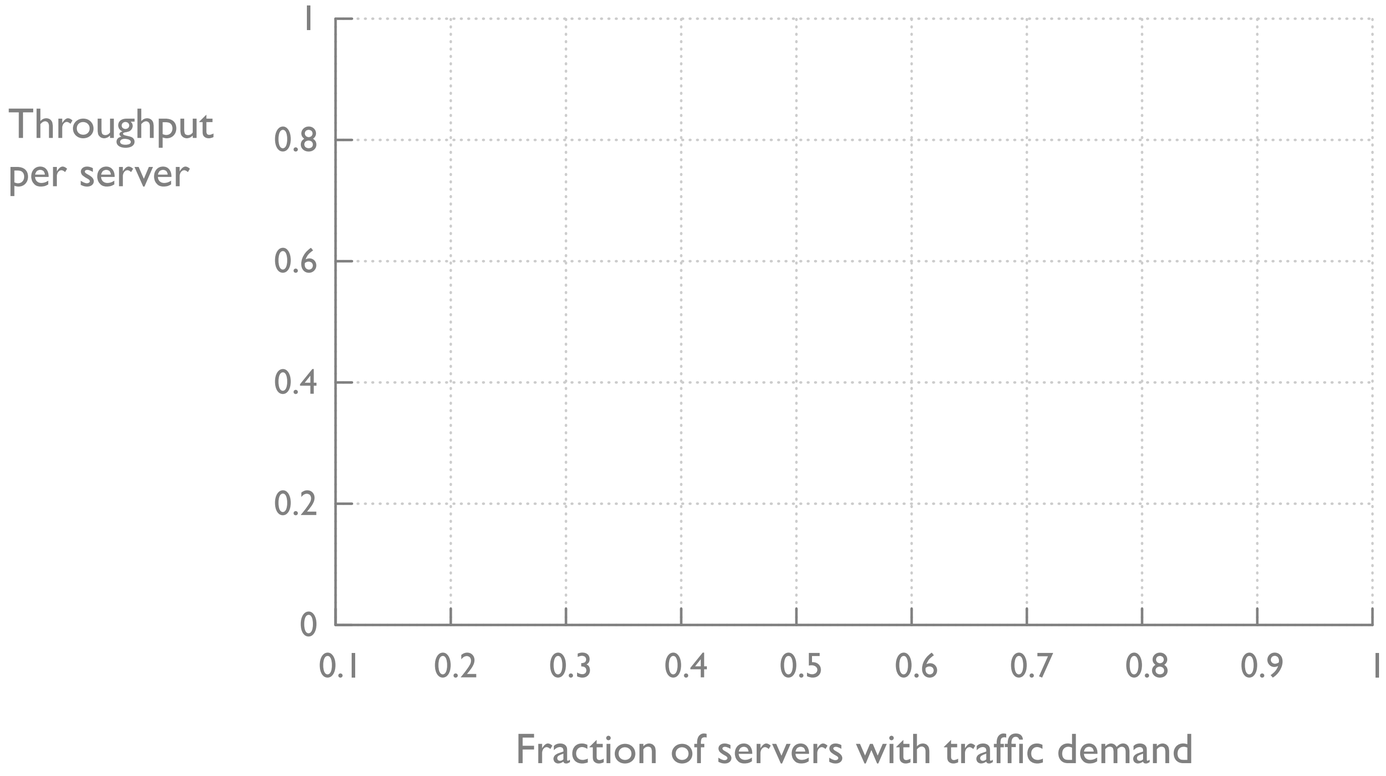


VS.

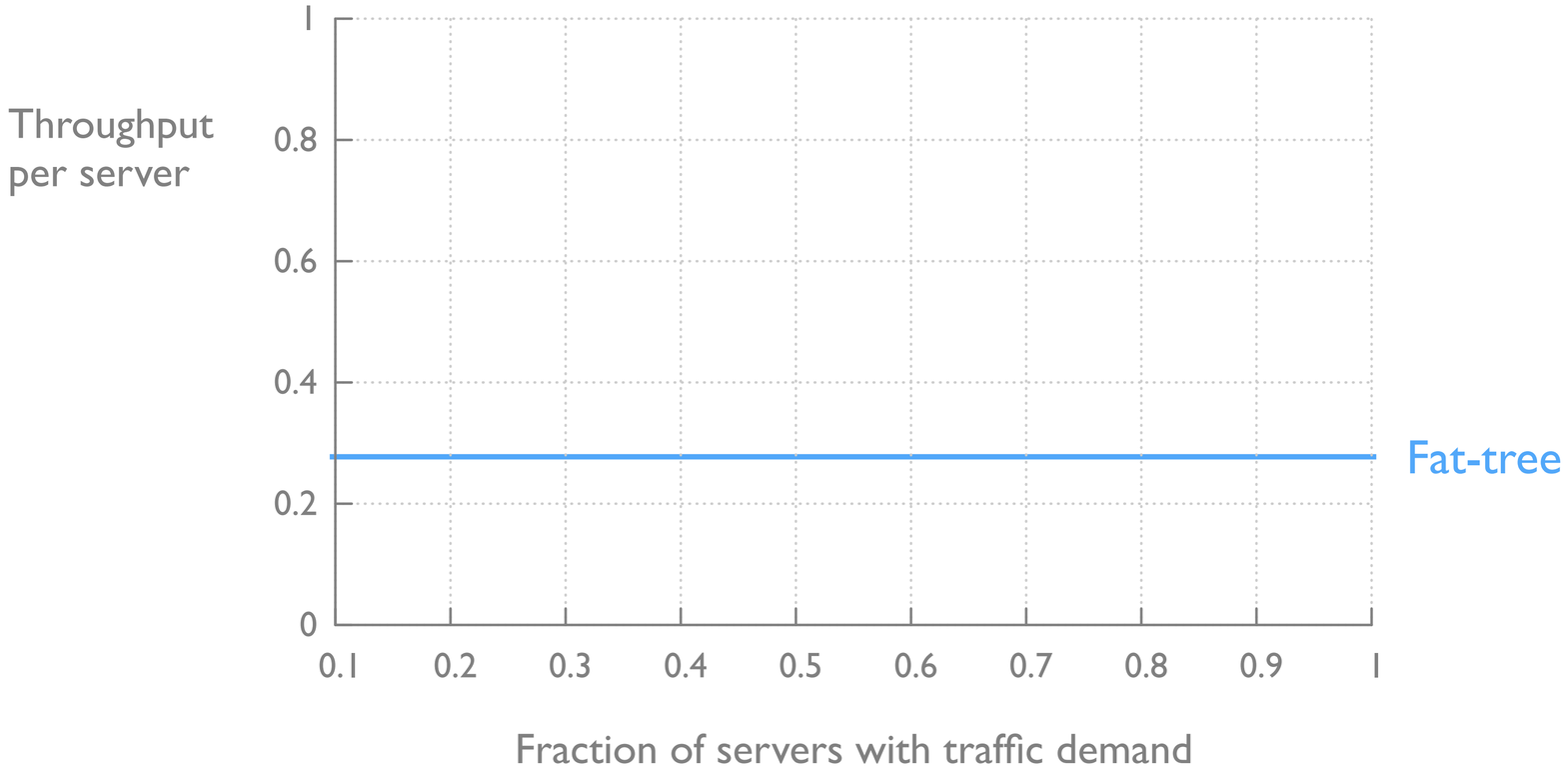


How **valuable** is the **ability to move links** around?

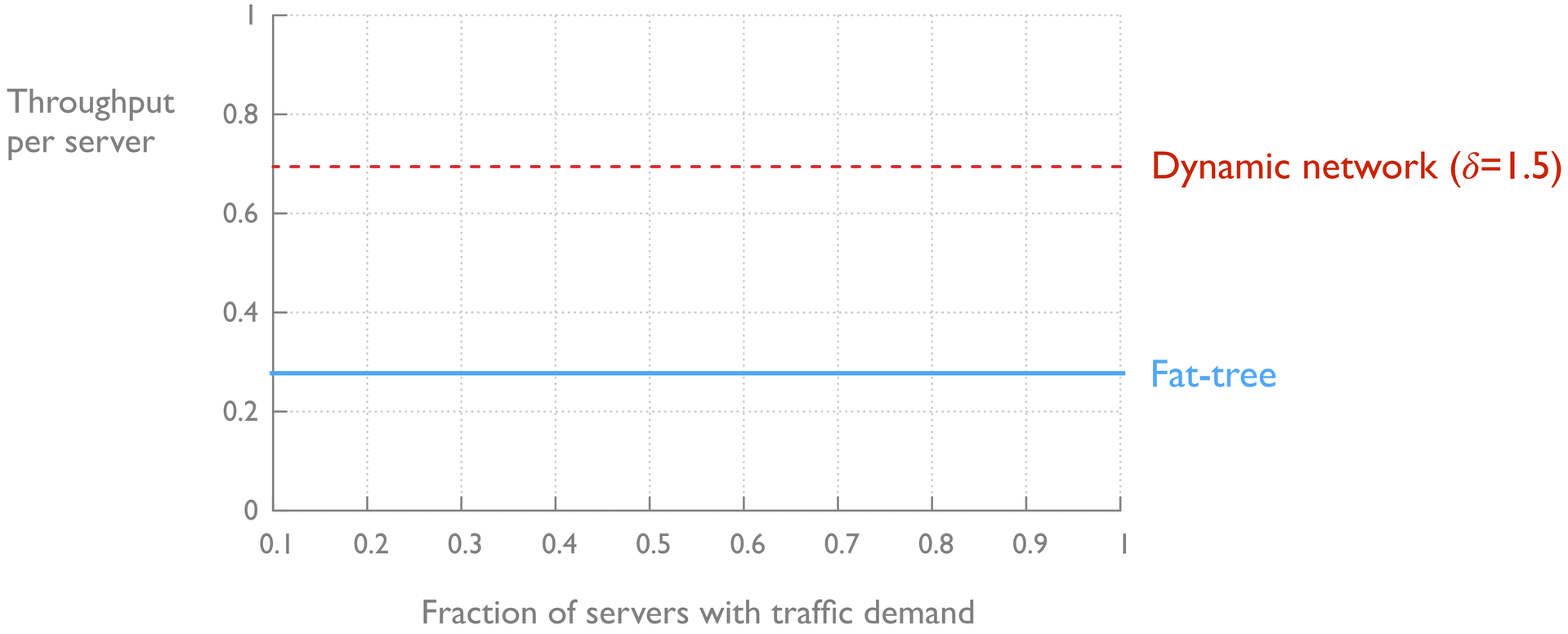
Optimal flow comparison



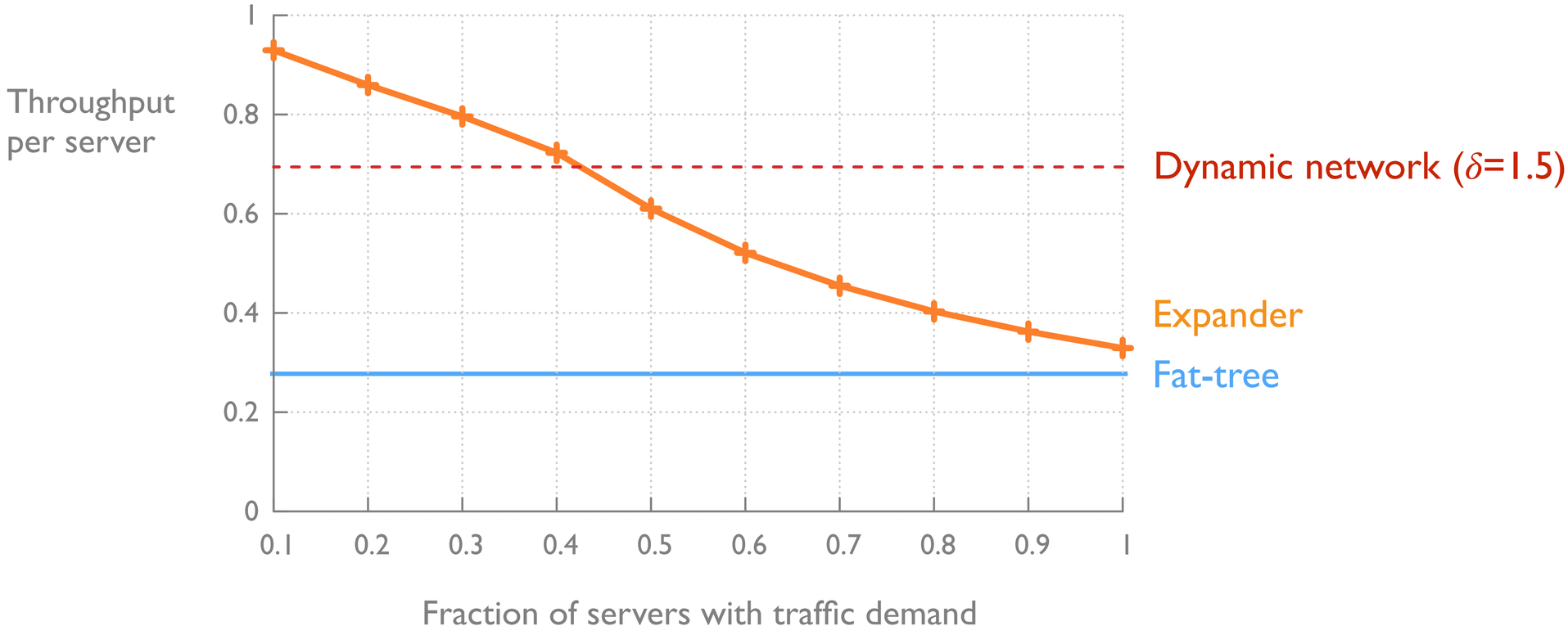
Baseline: oversubscribed fat-tree



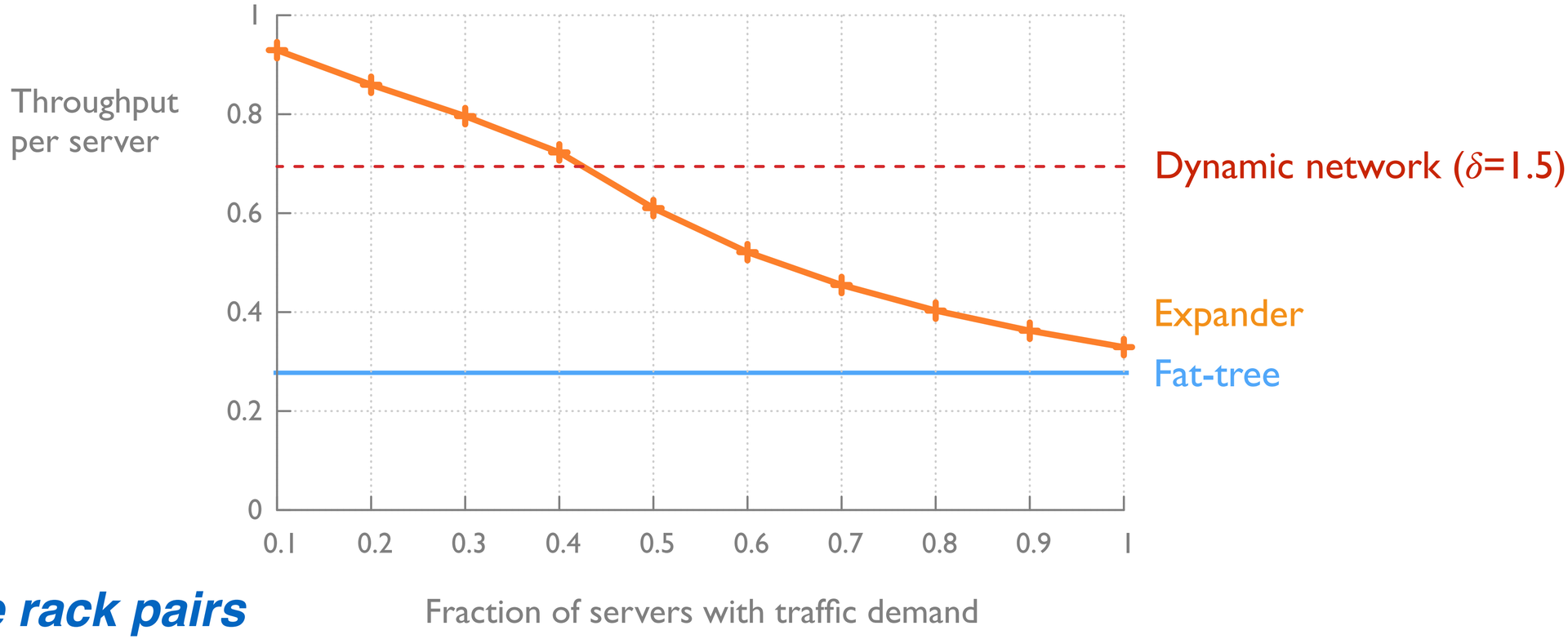
Indeed, dynamic networks can be better



… but so can static ones



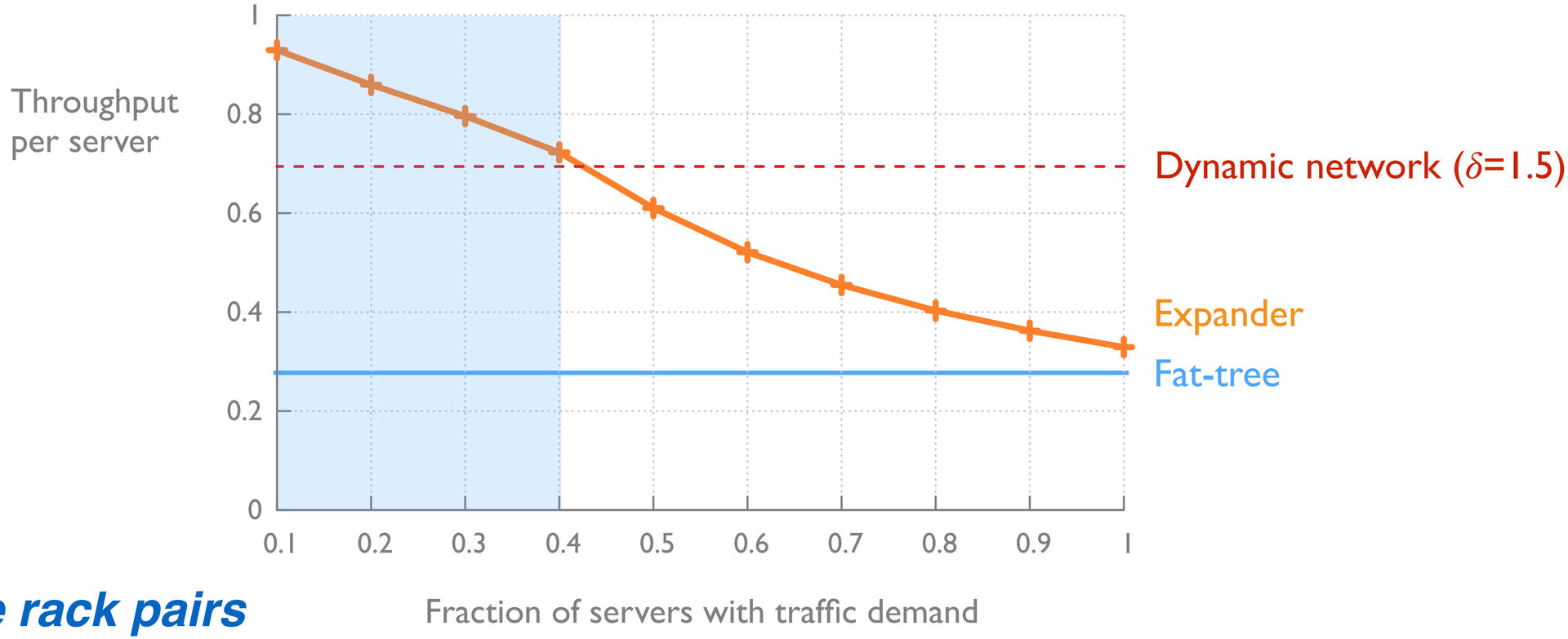
… especially in the regime of interest



"46-99% of the rack pairs exchange no traffic at all"

— Ghobadi et al., 2016

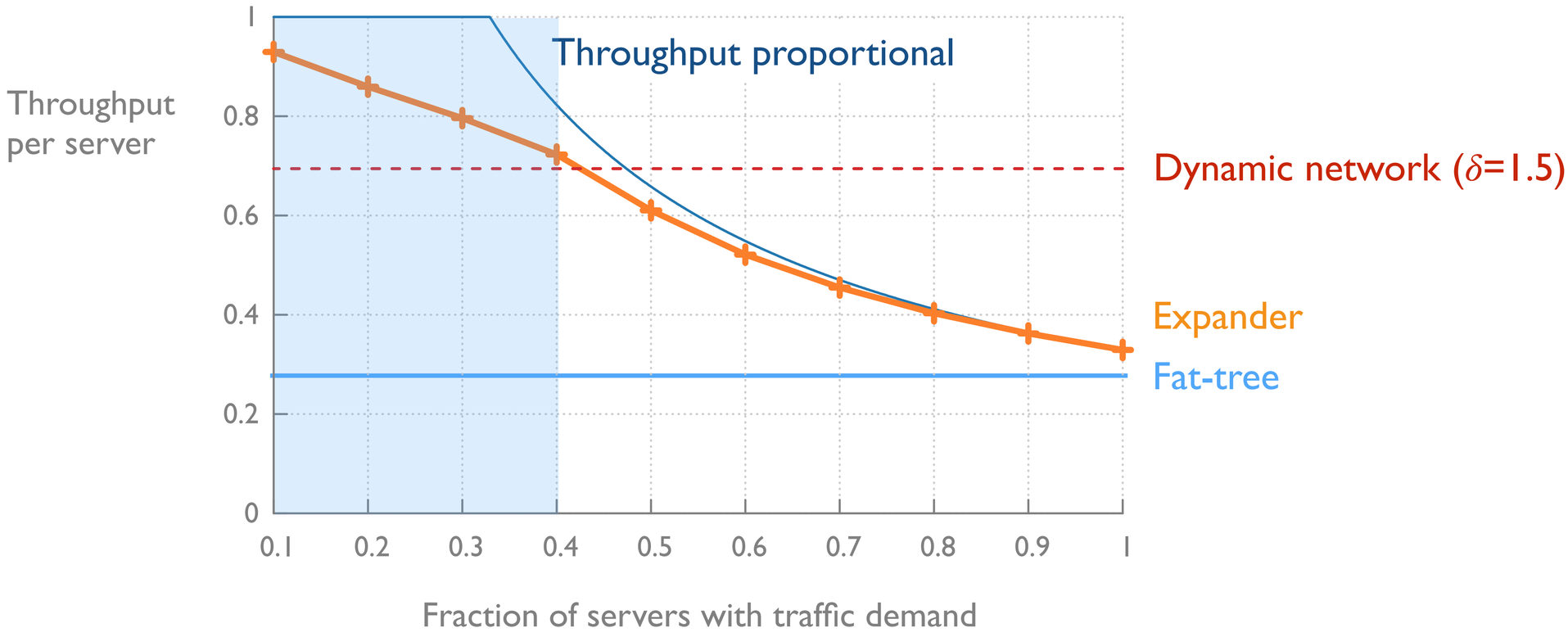
… especially in the regime of interest



"46-99% of the rack pairs exchange no traffic at all"

— Ghobadi et al., 2016

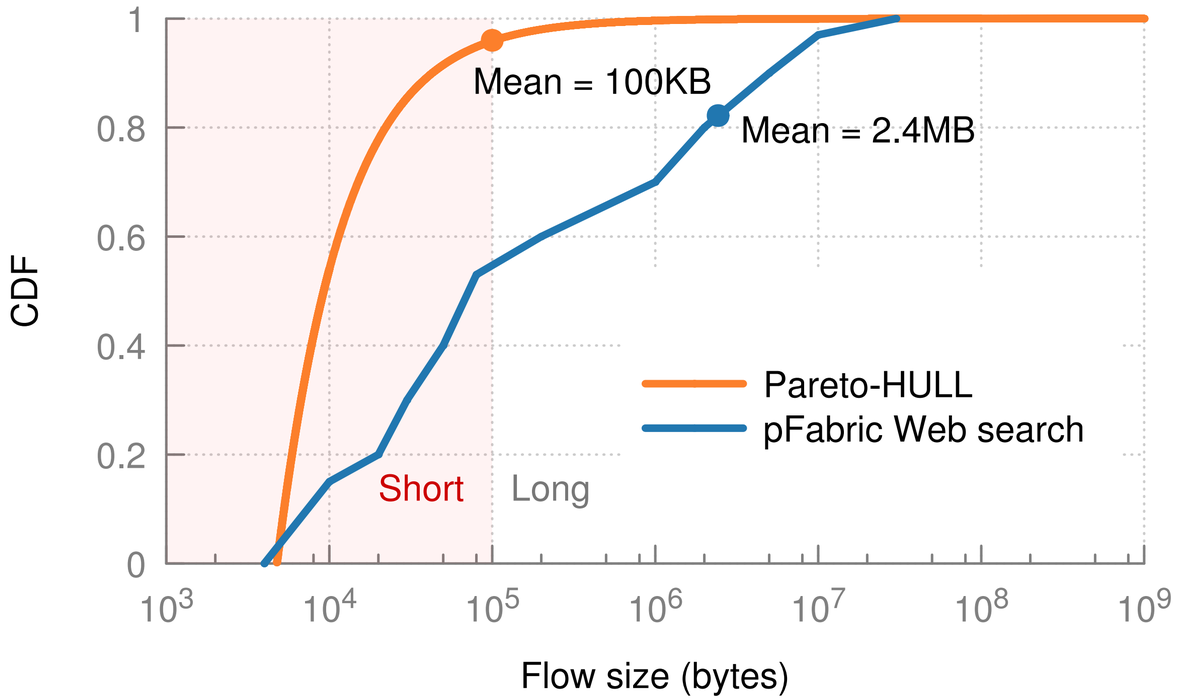
Not too far from proportionality!



Workloads

pFabric Web search (2.4MB mean)
Modelled after a real workload
Maximum flow size of 30MB

Pareto-HULL (100KB mean)
Pareto distributed
Highly skewed
Many short flows (<100 KB)
Few very large flows (max. 1GB)



… at a fixed arrival rate per second (λ)

Pareto-HULL : Alizadeh, Mohammad, et al. "Less is more: trading a little bandwidth for ultra-low latency in the data center." USENIX (2012).
pFabric Web search : Alizadeh, Mohammad, et al. "pfabric: Minimal near-optimal datacenter transport." SIGCOMM (2013).

Traffic scenarios

A2A(x): fractional all-to-all
Only the servers under x% of the ToRs communicate all-to-all

Permute(x): fractional random permutation
A random pairing of x% of the ToRs, of which in each pair all servers only communicate with the servers of the counterpart

ProjecToR
Empirical skewed traffic from a Microsoft cluster

Skew(x, y)
x fraction of ToRs has *y* probability of participating in a flow (rack-pair)
E.g. θ=4% of ToRs have ϕ=77% chance of participating in a flow

ProjecToR: Ghobadi, Monia, et al. "Projector: Agile reconfigurable data center interconnect." SIGCOMM *(*2016).

Topologies & Routing

Two topologies (k=16):

- Full fat-tree with n=320
- Xpander at with n=216 (67.5%)

… with 10 Gbps links

… both supporting ~1K servers

At servers:

- DCTCP
- Flowlets (change path upon exceeding gap)

Fat-tree:

- ECMP

Xpander:

- HYBRID

Introducing **HYBRID** routing

HYBRID routing:

- ECMP until # sent bytes > threshold Q
- After threshold Q, use valiant load balancing (VLB)

Advantages:

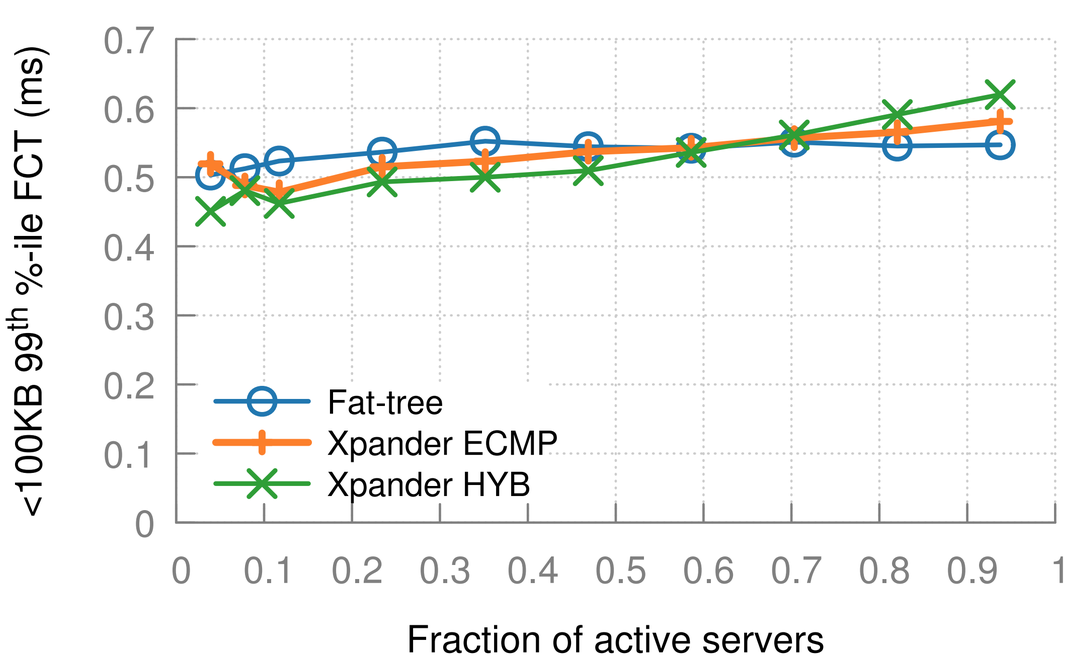
- Oblivious to the network congestion state
- Introduces little to no overhead in current switches

Experimental take-aways

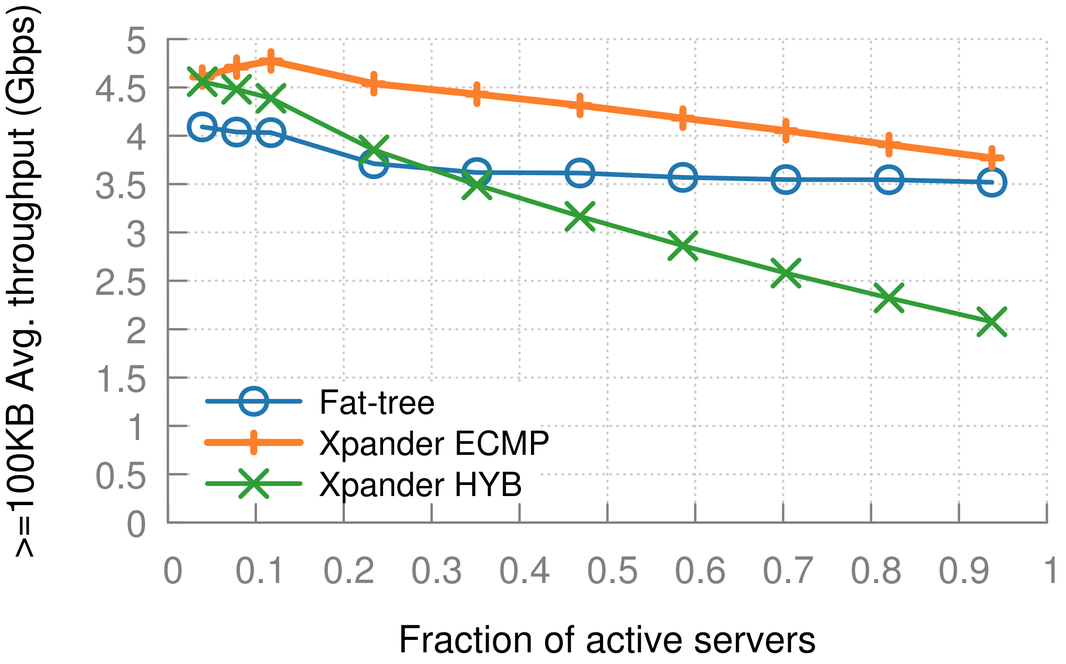
- Xpander achieves **comparable performance** to non-blocking fabrics…
- At **lower cost**: 2/3rds or less
- Matching the performance of dynamic topologies

A2A(x): fractional all-to-all (pFabric)

99th %-tile FCT for small flows
(lower is better)

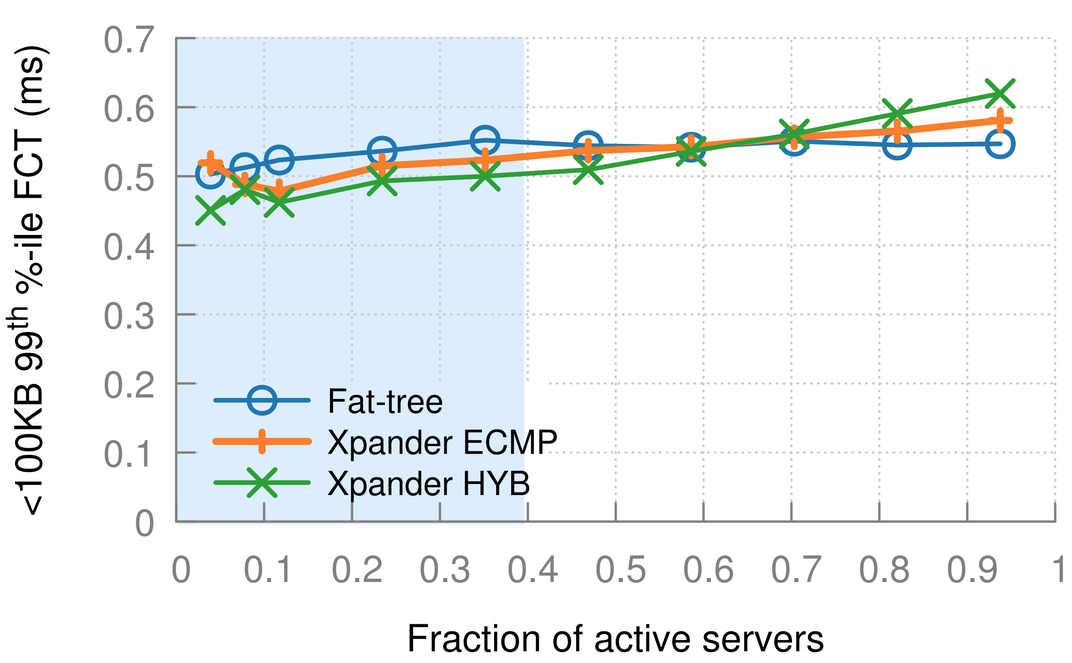


Average throughput for large flows
(higher is better)

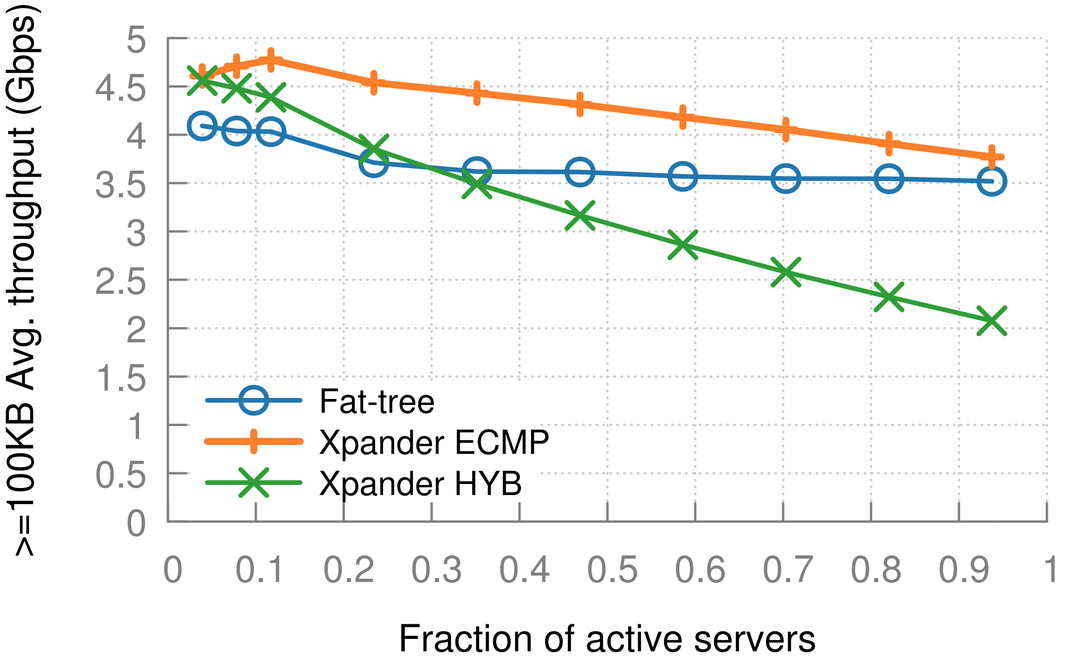


A2A(x): fractional all-to-all (pFabric)

99th %-tile FCT for small flows
(lower is better)

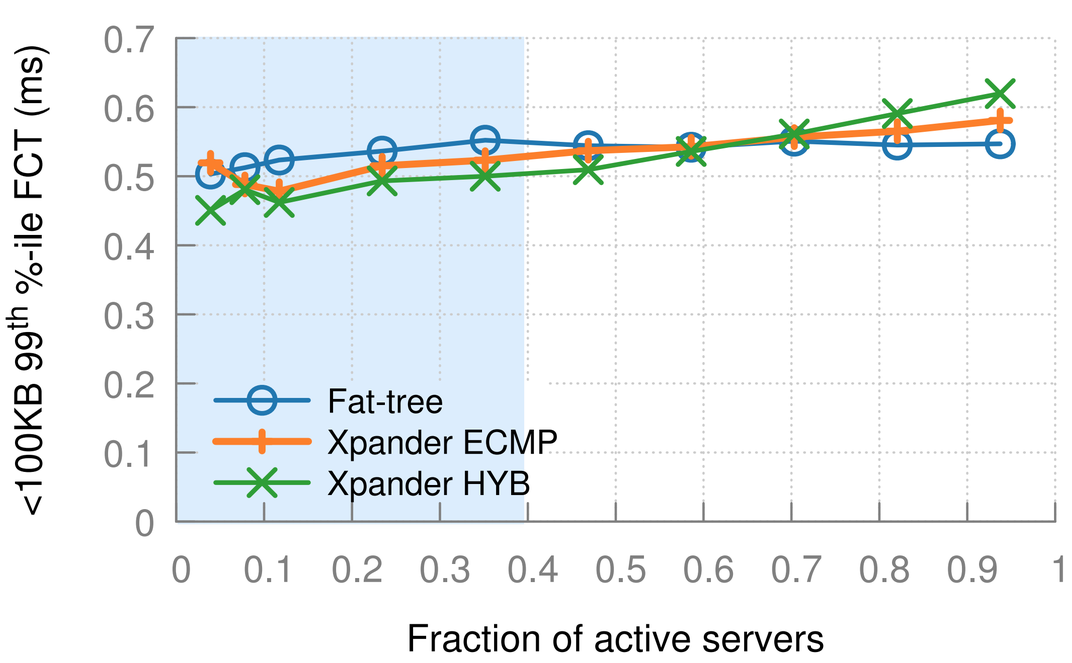


Average throughput for large flows
(higher is better)

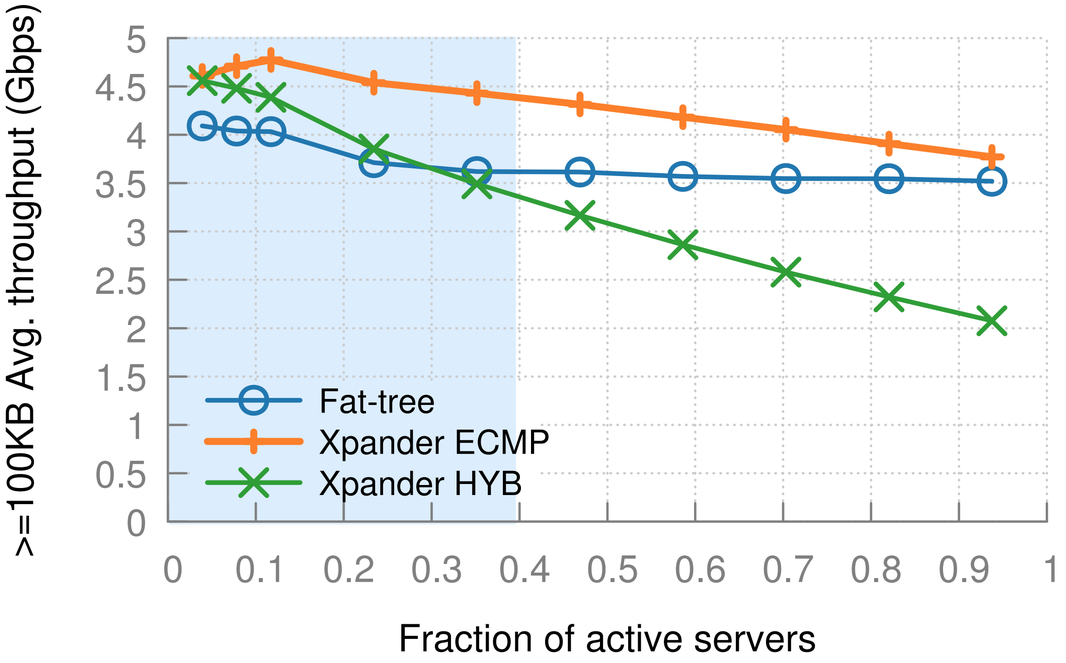


A2A(x): fractional all-to-all (pFabric)

99th %-tile FCT for small flows
(lower is better)

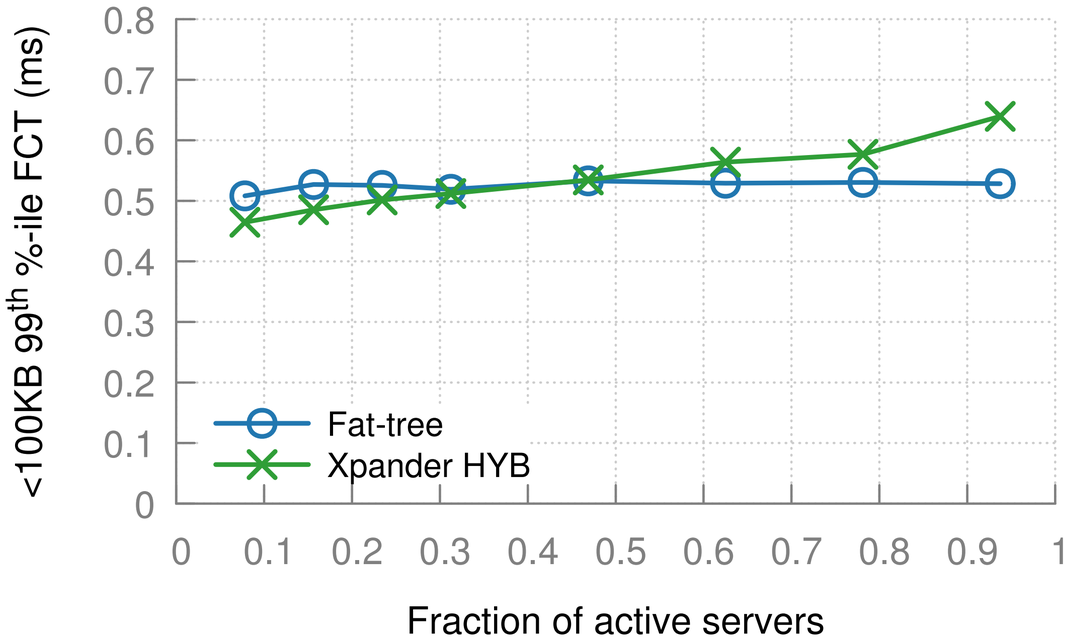


Average throughput for large flows
(higher is better)

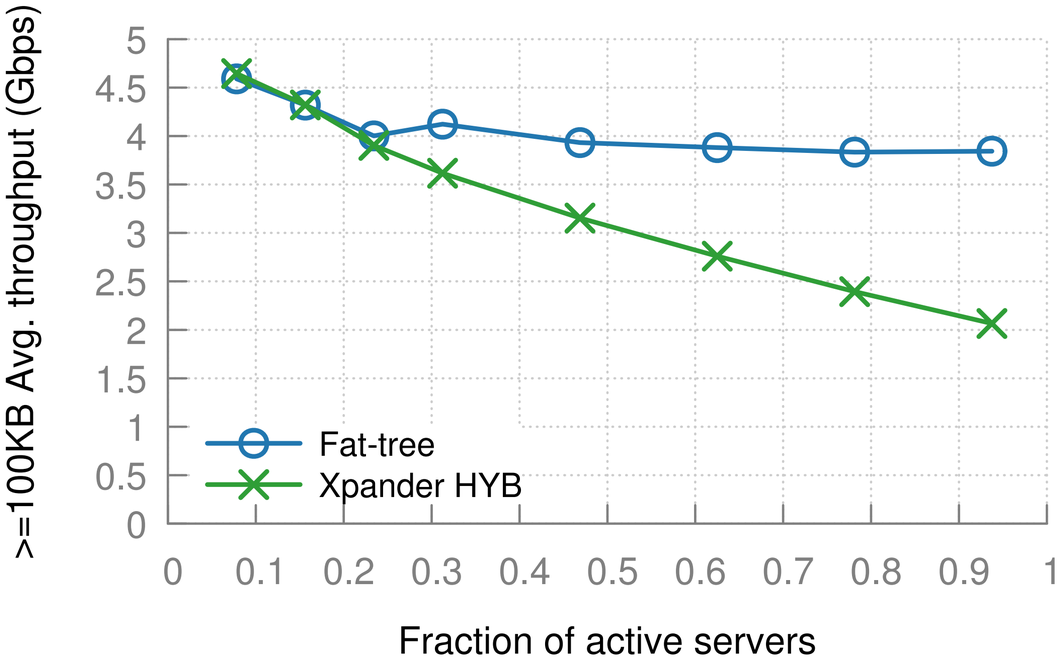


Permute(x): fractional random permutation (pFabric)

99th %-tile FCT for small flows
(lower is better)

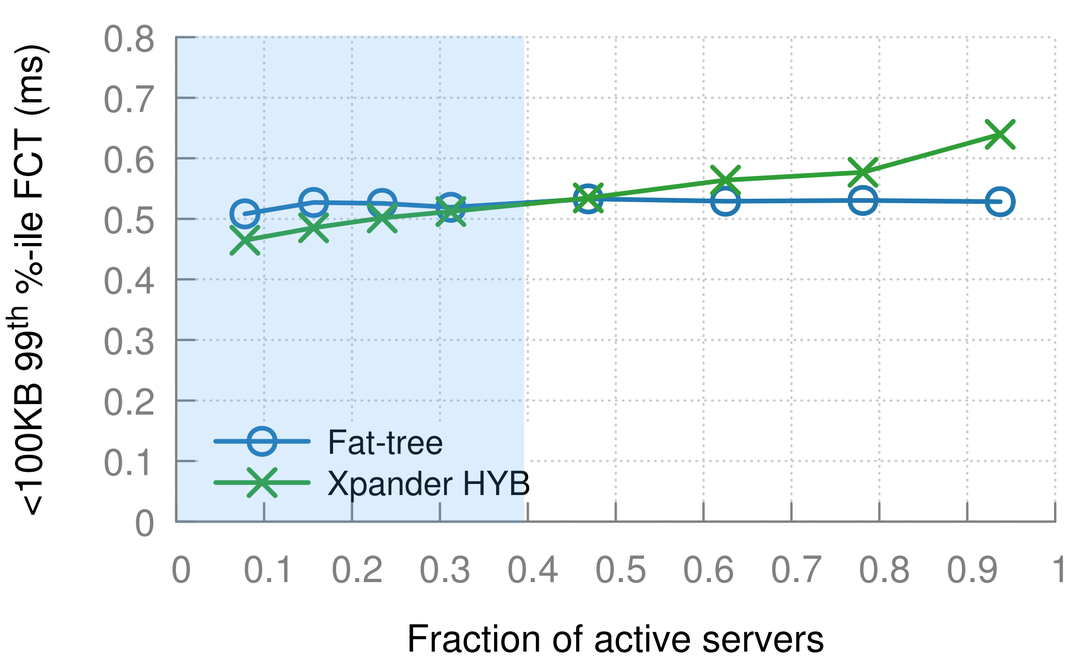


Average throughput for large flows
(higher is better)

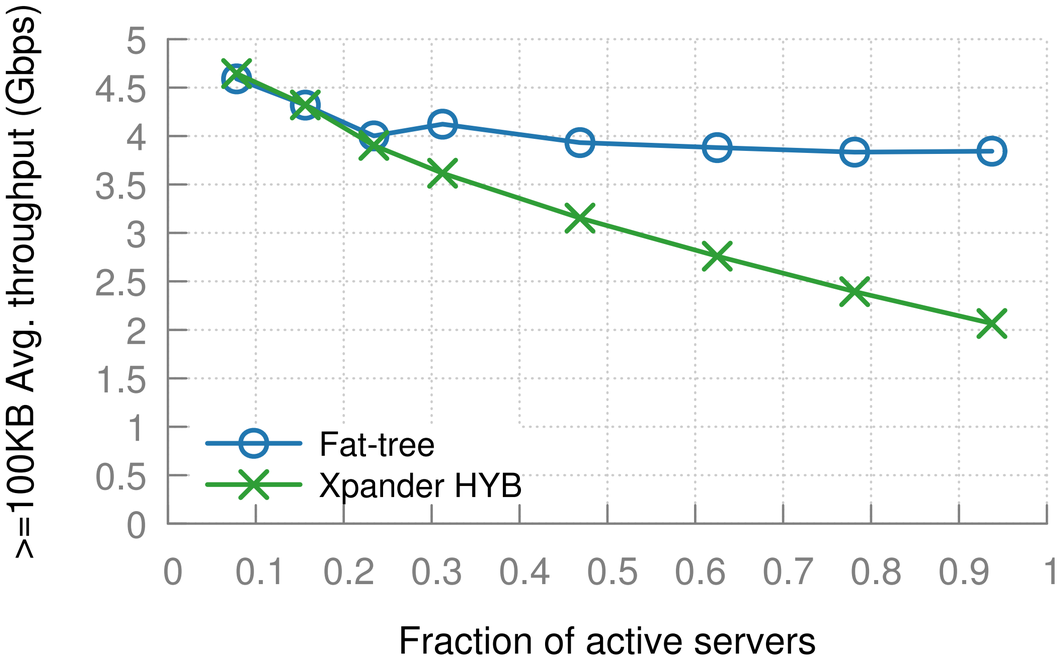


Permute(x): fractional random permutation (pFabric)

99th %-tile FCT for small flows
(lower is better)

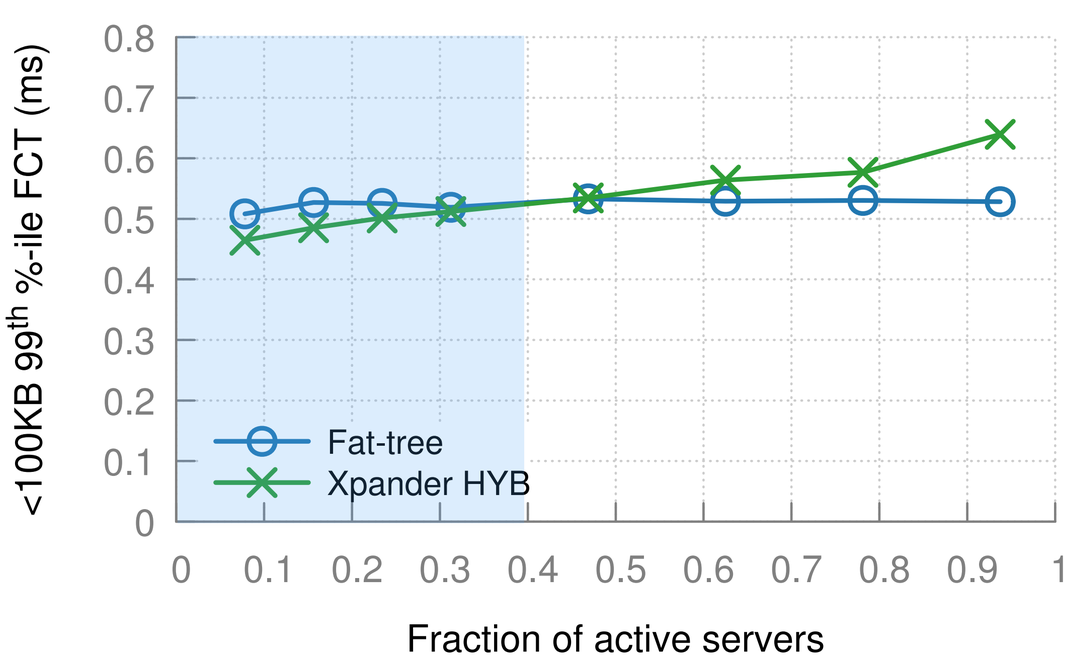


Average throughput for large flows
(higher is better)

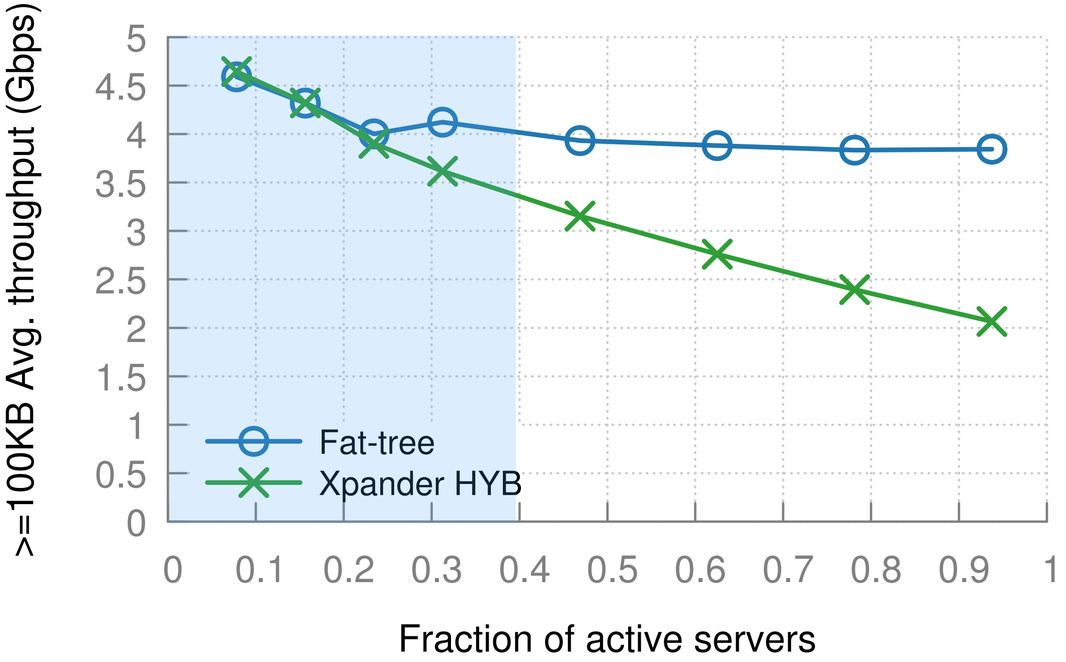


Permute(x): fractional random permutation (pFabric)

99th %-tile FCT for small flows
(lower is better)

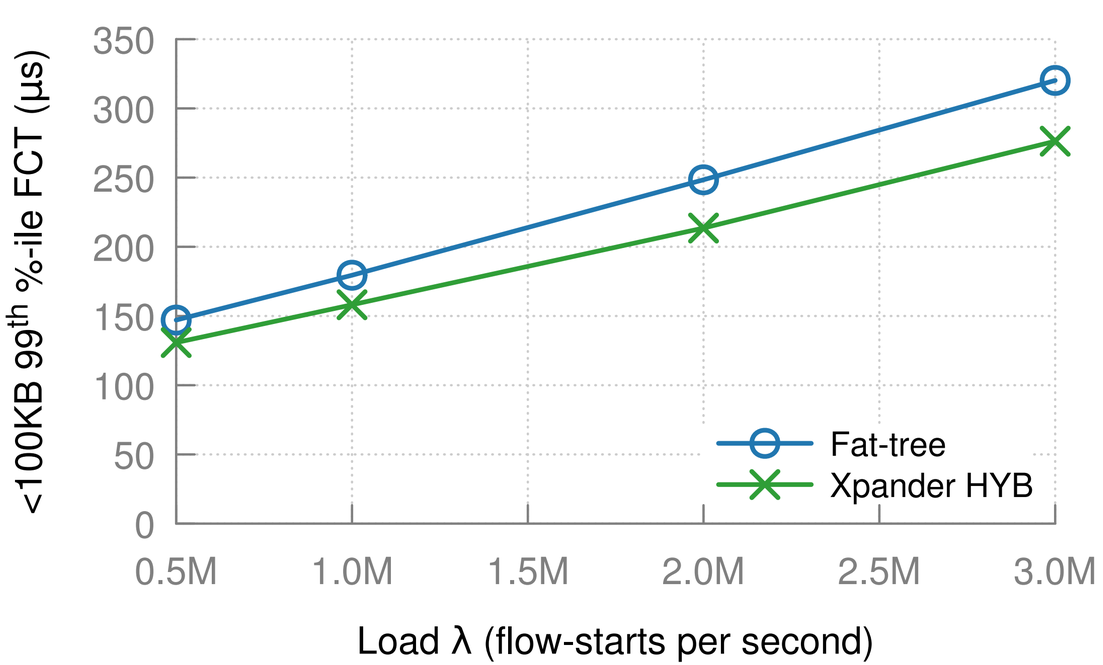


Average throughput for large flows
(higher is better)

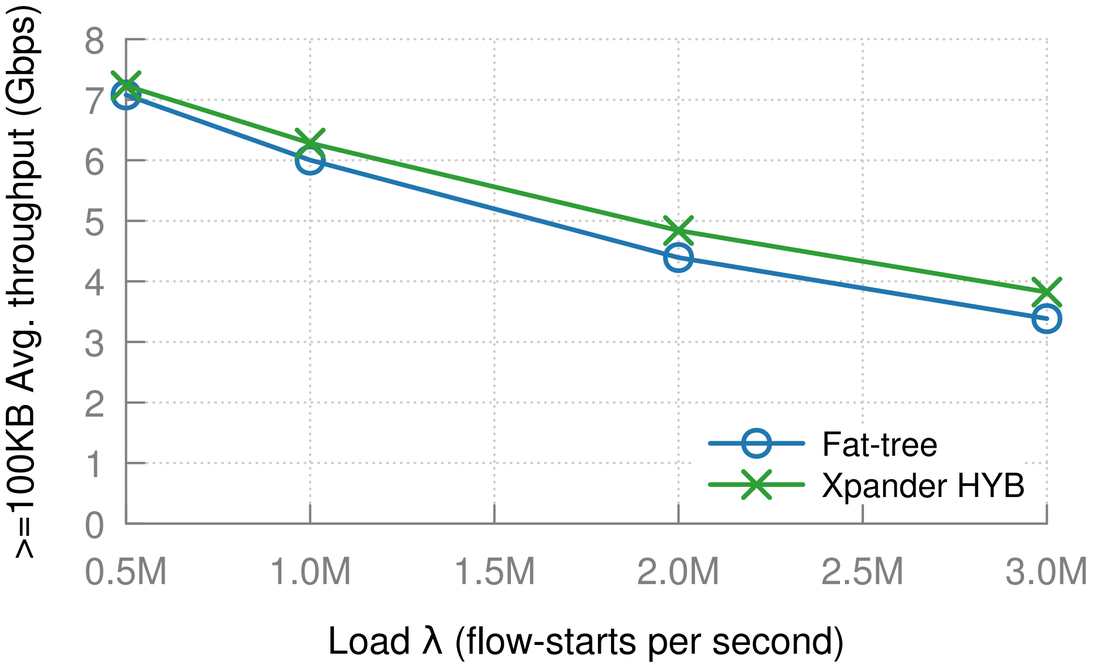


A2A(0.31) with many short flows (Pareto-HULL)

99th %-tile FCT for small flows
(lower is better)



Average throughput for large flows
(higher is better)

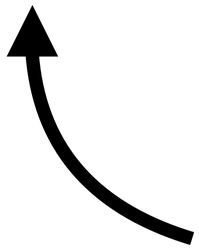


Comparing against ProjecToR

- Creating the same experiment as ProjecToR
- Same workload (pFabric)
- Same traffic scenario
- Same network sizes:

k=16 fat-tree: 320 switches

d=16, r=8 Xpander: 128 switches (40%)

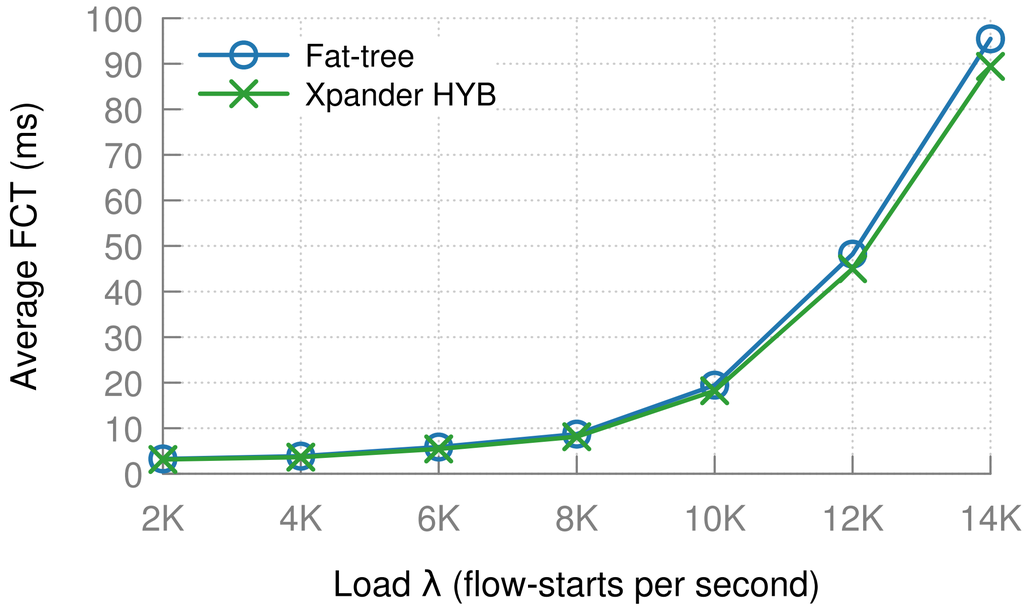


Static links

ProjecToR: same # of network ports but static

Average FCT
for all flows
(lower is better)

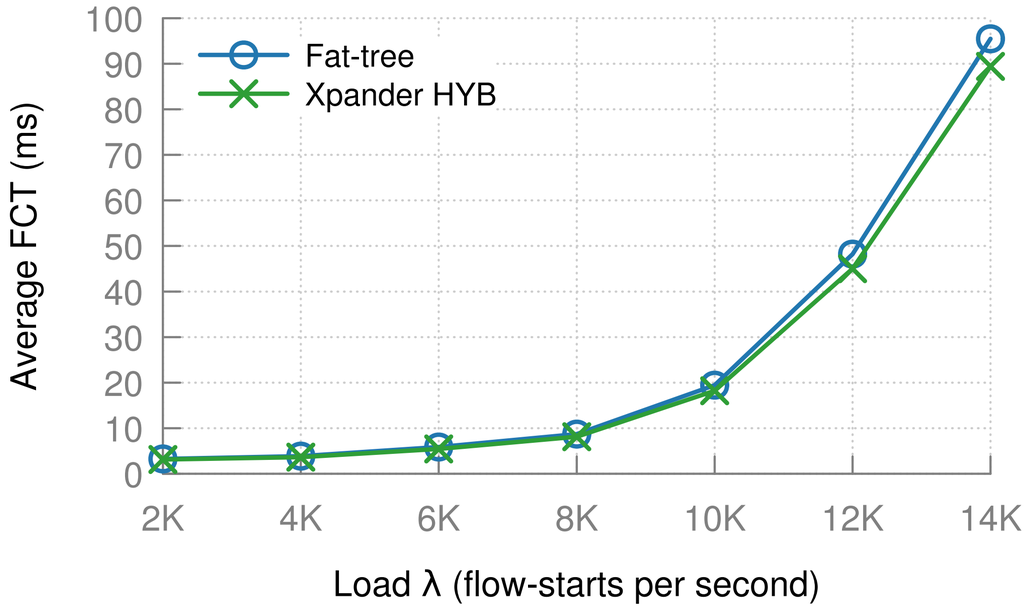
Empirical



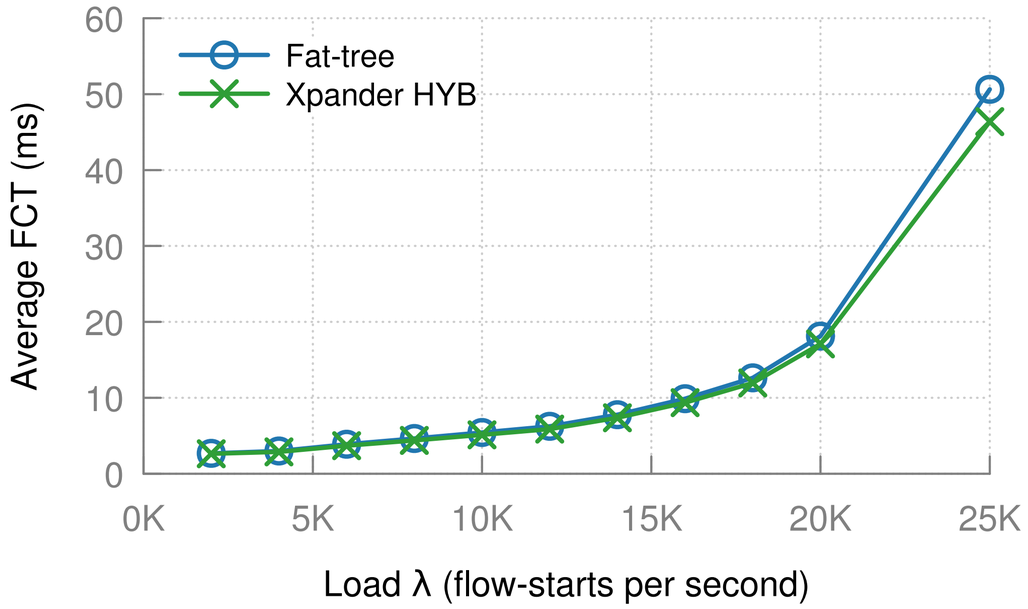
ProjecToR: same # of network ports but static

Average FCT for all flows
(lower is better)

Empirical



Skew (4%, 77%)



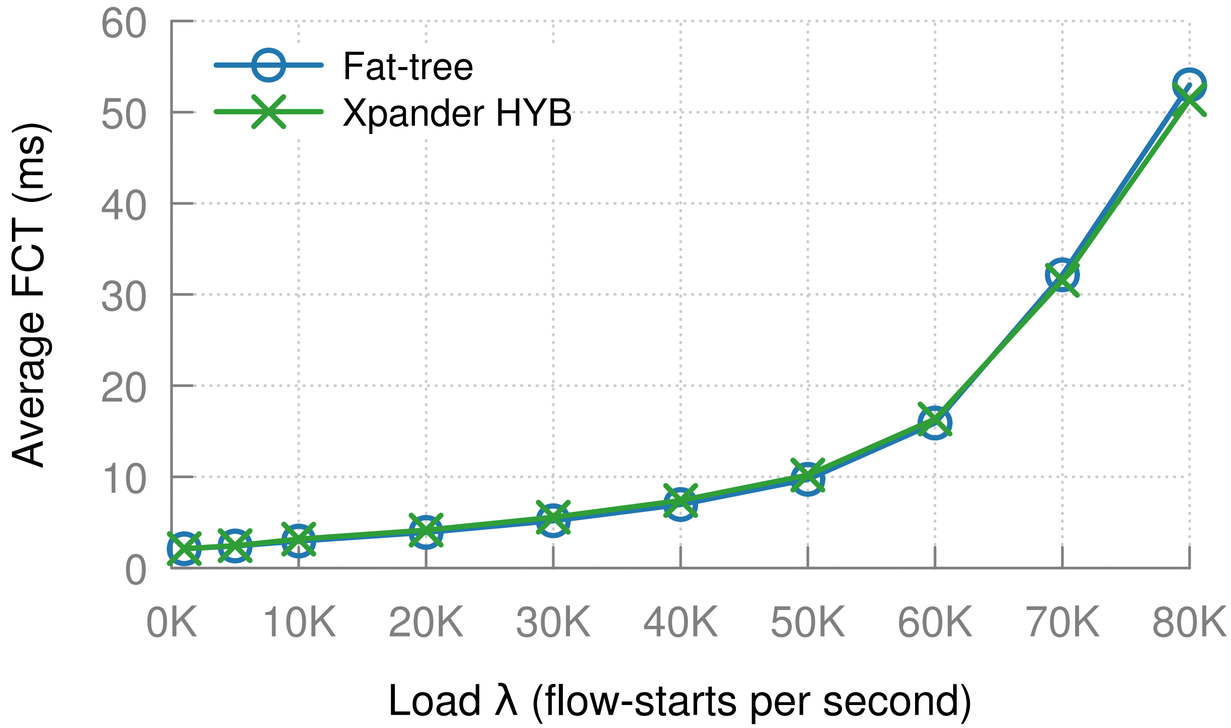
Skew(4%, 77%) using same equipment at larger scale

k=24 fat-tree
(720 switches)

d=13, r=11 Xpander
(322 switches = 45%)

… both supporting
~3.5k servers

Average FCT
for all flows
(lower is better)



cheaper expander + simple, practical routing

=

performance of full-bandwidth fat-tree

Expanders: the static topology benchmark

Demonstrating an advantage of dynamic topologies over static topologies requires…

- … comparing to expander-based static networks
- … at equal cost
- … using more expressive routing than ECMP
- … accounting for reconfiguration/buffering latency

All proposals to date don't hit this benchmark

Future work

A. Better (oblivious) routing schemes?

B. Adaptive routing?

C. Deployment?

Get in touch

My e-mail: simon.kassing [at] inf.ethz.ch

Code available: https://github.com/ndal-eth/netbench

ETH zürich

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM